UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Facoltà di Scienze Matematiche, Fisiche e Naturali

Dipartimento di Biotecnologie e Bioscienze

Dottorato di Ricerca in Biotecnologie Industriali

High-throughput bioinformatic approaches to study tumorigenesis in mammalian cells

PhD: Dr. Chiara BALESTRIERI

Supervisor: Dr. Ferdinando CHIARADONNA

Coordinator: Prof. Marco Ercole VANONI

2009 - 2010

List of Headings

Abstract

If we begin with certainties, we shall end in doubts,
but if we begin with doubts, and are patient in them, we shall end in certainties.
Francis Bacon

The analysis of transcriptional data has become increasingly populating in the last decade due to the advent of new high-throughput technologies in genome research, since the reported invention by the Pat Brown laboratory in 1995 (Schena et al., 1995) and by Affymetrix in 1996 (Lockhart et al., 1996). DNA microarray is a multiuse technology, in fact different technologies are employed to produce the microarray chips and different technical approaches are used for analyzing microarray data, ranging from statistical models of the decision process to machine-learning methods for identifying class predictors. The underlying technology is extremely complex. In fact, DNA microarrays generate large amounts of numerical data that should be analyzed effectively. Therefore the chosen of an appropriate analysis for DNA microarray experiments is the most important key to perform the assay and utilize the data correctly.

Genome wide expression profiling is a powerful tool for the investigation of novel gene ensembles in cellular mechanisms of health and disease. In fact, the DNA microarray expression analysis can be used to study complex multigenic diseases such as cancer. The great challenge in understanding the genetics of such disorders is the identification of susceptibility genes, which are genes that increase a person's risk of developing the disease. Decades of molecular genetics researches have shown that cancer is a heterogeneous cellular disorder caused by the deregulation of many interacting cellular pathways that converge to generate tumor formation and growth. Since the draft sequence of the human genome was published in 2001 (Lander et al., 2001) the Cancer Genome Anatomy Project index of tumor genes has classified more than 40000 genes directly or indirectly involved in one or more cancers (Strausberg, 2001; Strausberg

et al., 2000). The rapid accumulation of high-resolution cancer genetic data, now promises to enable far more comprehensive and unbiased inference of uncharacterized cancer genes linked to complex tumor traits such as metastasis and angiogenesis (Vogelstein and Kinzler, 2004).

During the last three years, I focused the attention on the analysis and the interpretation of GeneChip data, with the aim of setting up workflows useful to characterize different cellular physiological and pathological (i.e., cancer) conditions, to dissect the effects of nutrient perturbations on cell culture models, to interpret time-dependent gene expression fluctuations as well as to identify, by orthologous comparisons, phylogenetic conservation of promotorial regulative sequences and cancer cells signatures.

Taking into consideration that the development of efficient methods that facilitate the biological interpretation of these data is crucial, in this thesis the work has been focused on some new ideas and analytical methods in order to get an efficient identification of cancer regulatory mechanisms. In this regard my thesis work proposes the use of several approaches for analysis and interpretation of gene expression data, based on the integration of different types of related biological information and software tools for efficient data analysis.

The most important contribution of this thesis to the scientific community is the proposal of integrating different “omic” approaches for the study of systemic disease as cancer. It is worth pointing that the proceeding in this way requires gathering information from several fields, such as molecular biology, biochemistry, mathematic, informatics, statistic ect., which altogether provide fundamental knowledge to establish the contextualized study’s framework.

Introduction

1. What is bioinformatics?

Life sciences are currently at the center of informational devolution. Dramatic changes are begin registered as a consequence of the development of techniques and tools that allow the collection of biological information at an unprecedented level of detail and in extremely large quantities. The human genome project is a compelling example. The project was developed from an idea discussed at scientific meetings in 1984 and 1985 and a pilot project was begun by the Department Of Energy (DOE) in 1986 (Waterston et al., 2002). Initially, the plan to map the human genome was considered extremely ambitious, on the border of feasibility, but the entire genome was mapped in less than 3 years, at a much lower cost than initially expected. The nature and amount of information now available open directions of research that were once in the realm science fiction. Pharmacogenomics (Roses, 2000), diagnostic (Ross et al., 2000; Wellmann et al., 2000) and drug target identification (Marton et al., 1998) are just few of many areas that have the potential to use this information to change dramatically the scientific landscape of life sciences.

Bioinformatics is an emerging discipline situated at the interface between computer science and biological sciences, such as molecular biology and genetics. Initially, the term bioinformatics was used to denote very specific tasks such as the activities related to the storage of data of biological nature in databases. As the field evolved, the term has started to encompass also algorithms and techniques used in the context of biological problems. Today the field of bioinformatics supports a broad spectrum of research including determination of the significance of vast biological data, provides the expertise to organize it, and develops practical computational tools needed to mine the data for the new information. The definition submitted to the Oxford English Dictionary, represents as well the no clear universally definition of bioinformatics.

> (Molecular) bio-informatics: bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical chemistry) and applying "*informatics techniques*" (derived from disciplines such as applied math, computer science and statistics) to *understand* and *organize* the *information* associated with these molecules, on a *large scale*. In short, bioinformatics is a management information system for molecular biology and has many *practical applications*.

In other word, the bioinformatics is a discipline that stores all the information and turns it into understandable trends and facts that users can readily understand.

2. DNA Microarray Technology

Although all of the cells in the human body contain identical genetic material, the same genes are not active in every cell. Studying which genes are active and which are inactive in different cell types helps scientists to understand both how these cells function normally and how they are affected when various genes do not perform properly. In the past, scientists have only been able to conduct these genetic analyses on a few genes at once. DNA microarray is a technology which enables the researchers to investigate and address issues which were once thought to be non traceable as the expression of many genes in a single reaction quickly and in an efficient manner. With the development of DNA microarray technology, however, scientists can now examine how active thousands of genes in a single experiment are (Schena et al., 1995), in order to understand, for instance, the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body. DNA microarrays are a significant advance because they contain a very large number of genes and because of their small size. Therefore, DNA microarrays are useful when there is the need to survey a large number of genes quickly or when the study sample is small. These technologies may be used to assay gene expression within a single sample or to compare gene expression in 2 different cell types or tissue samples, such as in healthy and diseased tissues. Because a DNA microarray can be used to examine the expression of hundreds or thousands of genes at once, it promises to revolutionize the way gene expression is examined.

2.1 The principle core

Array technology was in use as early as the 1980s (Augenlicht et al., 1984; Augenlicht et al., 1987) but did not come into prominence until the mid 1990s when cDNA microarrays emerged as an exciting new biomolecu-

lar tool capable of probing the entire transcriptome of the cell (DeRisi et al., 1996; Lockhart et al., 1996). As mentioned previously, the DNA microarrays technology (sometimes called DNA chips) is an assay for quantifying the types and amounts of mRNA transcripts present in a collection of cells. All microarray experiments rely on the core principle that the number of mRNA molecules derived from transcription of a given gene is an approximate estimation of its level of expression.

Microarrays chip are microscope slides on which strands of polynucleotides have been attached in specified positions. We refer to the polynucleotides immobilized on the solid surface as *probes*. The probes consist either of cDNA printed on the surface or shorter *oligonucleotides* synthesized or deposited on the surface. The labeled targets bind by hybridization to the probes on the array with which they share sufficient sequence complementarity. The property of nucleic acid sequences of specifically build hydrogen bonds with the complementary nucleotide base pairs is the core technique principle of microarrays. A high number of complementary base pairs in a nucleotide sequence mean tighter non-covalent binding between the two strands. After washing off of non-specific bound sequences, only strongly paired strands will remain hybridized. Fluorescently labeled target sequences bind to a probe sequence that generate a signal that depends on the strength of the hybridization determined by the number of paired bases. Total strength of the signal, from a spot, depends upon the amount of target sample binding to the probes present on that spot.

2.2 GeneChips technologies

The microarray platforms used to generate the raw data, known as the image file, are different and are closely related to the nature of study and consequently the experimental design. These various platforms were optimized and validated during their development to maximize the accuracy of data. Each of them has a demonstrated efficiency with respect to its signal

dynamic range, discrimination power, reproducibility of raw data and fold change or expression level values. In this section the most commonly marketed *GeneChip platforms*, used in every work of this thesis, will be briefly presented, for a detailed description and comparison of different platforms refer to (Hardiman, 2004).

The new microarray technology is *in situ-synthesized oligonucleotide* arrays that use a photolithographic technique. Affymetrix pioneered this field, and, consequently, their GeneChips have gained increasing acceptance as the optimal method for determining transcriptional profiles with high level of reproducibility. The GeneChip platform consists of short single-stranded DNA segments, oligonucleotides or oligos, which are built by chemical synthesis (Chee et al., 1996) and in this case the array is not a glass slide (as traditional microarray), but a silicon chip (Fodor et al., 1991). The oligonucleotides at all locations on the chip, are synthesized in parallel. Light-directed DNA synthesis is employed to construct high-density array using a combination of two techniques, photolithography and solid-phase DNA synthesis. Synthetic linkers containing photochemically removable protecting groups are attached to silicon substrate. Light is subsequently directed through a photolithographic mask to specific areas on the chip surface, producing localized photodeprotection. Chemical building blocks are incubated with the surface, and chemical coupling occurs at those sites that have been illuminated in the preceding step. The subsequent step requires light to be directed to different regions of the substrate using new masks and the chemical cycle is repeated. Arbitrary polynucleotides can thus be synthesized in a highly specific manner at defined locations (Hardiman, 2004). GeneChips are designed in silico, each gene target is probed by a number of distinct probes (10-25) collectively termed *probe set*. Some probes within a set are multiple independent oligonucleotides, in other word they hybridize to different regions of the same RNA (see Figure 1).

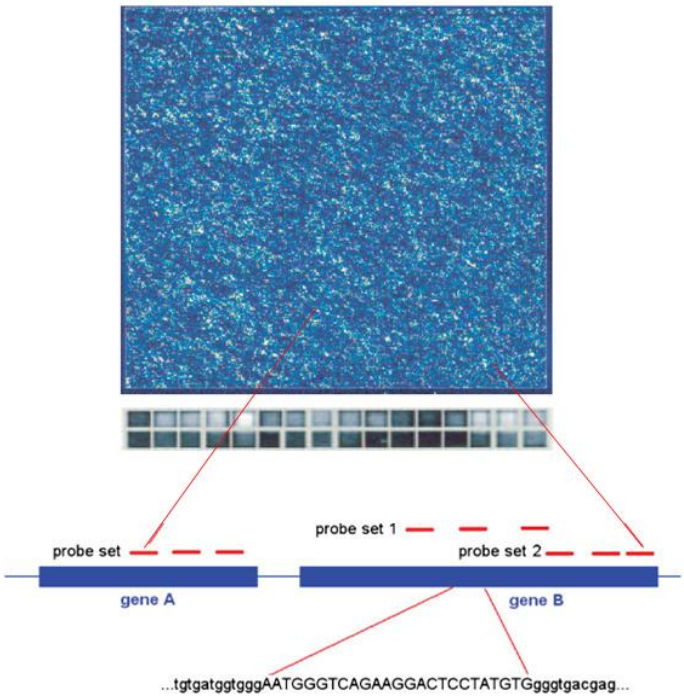

Figure 1: ***Example of image obtained by hybridizing the Affymetrix HG-U133 GeneChip.*** *Each gene is measured with one or several probe sets. Each probe set contains 11 probes which are 25 mer short oligonucleotides designed to match selected segments of exonic sequences of the gene, according to available data of gene sequences and annotations at the time the chip was designed (photolithography and in situ combinatorial chemistry-based synthesis). Accompanying each PM probe is a MM probe that is different from the PM probe only at the center base of the 25 mer. The probe pairs (the PM probes and their corresponding MM probes) are arranged on the chip at randomized locations to avoid systematic bias due to the chip layout. The basic unit in expression profiles obtained with such chips is probe set. For a gene with a single probe set on the chip, like gene A, the expression level of the probe set is the observed expression level of the gene. Genes with multiple probe sets, like gene B, will have multiple (not-necessarily consistent) observations of expression levels in the profile. A single-color labeling strategy was employed. The array contained ~ 45000 probes. The feature size is 18µm.*

A microarray experiment consists of the following components: a set of *probes*, an *array* on which these probes are immobilized at specified locations, a *sample* containing a complex mixture of labeled biomolecules that

can bind to the probes, and a *detector* that is able to measure the spatially resolved distribution of label after it has bound to the array.

The expense of fabrication and frequency of sequence errors for GeneChips increase with the length of the oligonucleotide probes employed, therefore relatively short 25 mer (25 bases in length) oligonucleotides are generally used. In order to obtain sufficient binding strength from 25 mer oligonucleotides, the hybridization conditions must be made less strict than for cDNA arrays or longer spotted oligonucleotide arrays. Consequently, substantial cross hybridization is possible. Probes are designed within 500 base pairs of the 3' end of each gene to hybridize uniquely in the same, predetermined hybridization conditions. Also an additional level of redundancy comes from the use of mismatch (MM) control probes identical to the perfect match (PM). In fact on high-density expression microarrays, a gene is usually interrogated using probes that either perfectly match the sequence in a segment of the target gene (PM probes), or contain a single mismatched nucleotide in the middle position of the corresponding perfectly matched probe (MM probes). The MM probes serve as controls for specific hybridization and they facilitate the direct subtraction of background and cross-hybridization signals (in order to minimizing cross-hybridization effects). Some housekeeping genes are represented as three probe sets, one set designed to the 5' end of the gene, the second set to the middle of the gene and the third to the 3'end. They serve as controls for the quality of the hybridized RNA. In addition to species-specific genes, some spiked-in control probe sets are introduced to facilitate the control of the hybridization. Biotinylated cRNA derived from a biological sample is hybridized onto the microarray, after allowing sufficient time for the hybridization, the excess sample is washed off the solid surface, stained and scanned for fluorescence at a single wavelength. Finally, the arrays are scanned, images are acquired and CEL files generated, which are then used for data analysis. At that point, each

probe on the microarray should be bound to a quantity of labeled target that is proportional to the level of expression of the gene represented by probe sets. Figure 1 shows an example of the fluorescence intensities on the chip recorded by an imaging device.

There are three principal differences between the Affymetrix GeneChip system and "traditional" cDNA microarrays in the study of gene expression. First, instead of hybridizing two RNAs labeled with different fluorophores competitively on one cDNA microarray, a single RNA is hybridized on the array in the Affymetrix system, and the comparisons are then made computationally. Second, GeneChip is not a competitive hybridization method and, in order to compare two samples, two separate microarrays are required. Third, in the Affymetrix arrays each gene is represented as a probe set of 10-25 oligonucleotide pairs instead of one full length or partial cDNA clone.

2.3 Bioinformatics work

A general approach to perform gene expression profiling experiments is indicated as a work flow diagram in Figure 2 (Gibson, 2003) and includes multiple steps, each of which created several specific bioinformatics challenges. The first four steps can be grouped together and define is the so-called "*Experimental Design*", then there are two levels of statistical analysis: the former constitutes the low-level investigation, that provides quality experiment and the removal of background noise while the latter, high-level analysis, includes the development of methods designed to answer to biological questions, in other words, the purpose is the extrapolation of the information.

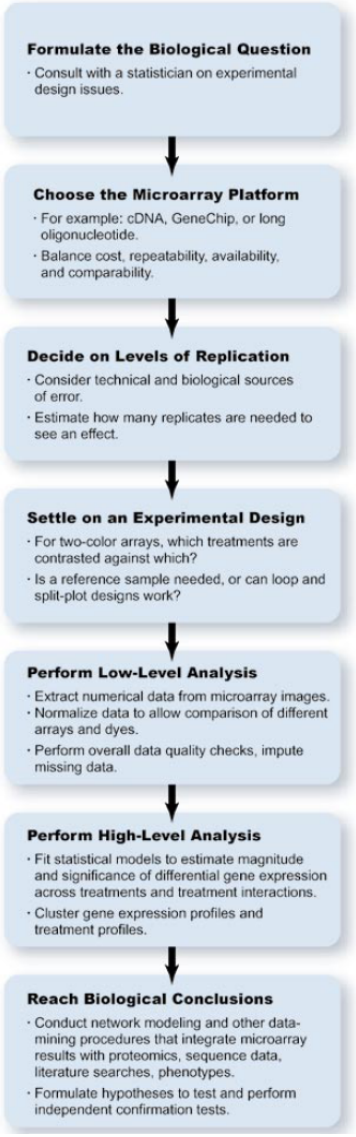

Figure 2: ***Example of gene expression profiling workflow*** *(Gibson, 2003).*

Experimental design

Due to the biological complexity of gene expression, microarray have multiple sources of variation, and experimental plans should ensure that effects of interest are not misunderstood with ancillary effects. The simplest microarray experiment looks for changes in gene expression across a single

factor of interest (*class comparison analysis*). Class comparison analysis focuses on determining whether gene expression profiles differ among samples selected from predefined classes and identifying which genes are differentially expressed among classes. For example, the class may represent different tissue types, the same tissue under different experimental conditions, or the same tissue type for different class of individuals. In cancer studies, the class often represents distinct categories of tumors differing with regard to stage, primary site, genetic mutations present, or with regard to response to therapy, the specimens may represent tissue taken before or after treatment or experimental intervention. In other word, the first step in choosing a good design is to identify which effects might possibly contribute to variation in the data between classes.

The initial task is to define the objectives of the experiment. Each experimental design should optimize the chances of answering to a key hypothesis. There is a natural temptation to test all of the interesting questions in a single experiment, but this approach is dangerous, as overly complex experiments may be un-testable, meaning the data are not statistically powerful enough to answer to all questions. In practice, this is the direct result of too few replicates or too little experimental controls.

There are three main elements to consider when designing a microarray experiment. First, replication of the biological samples is essential for drawing conclusions from the experiment. Second, technical replicates (two RNA samples obtained from each experimental unit) help to ensure precision and allow for testing differences within treatment groups. The technical replicates may be two independent RNA extractions or two aliquots of the same extraction. Third, spots of each cDNA clone or oligonucleotide are present as replicates (at least duplicates) on the microarray slide, to provide a measure of technical precision in each hybridization. Repeated spotting of the same clone on an array increases

precision (Lee et al., 2000; Lipshutz et al., 1999). It is critical that information about the sample preparation and handling is discussed, in order to help to identify the independent units in the experiment and to avoid inflated estimates of statistical significance (Churchill, 2002).

Statistical analysis: low-level investigation

Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. To properly compare summary measures of expression in terms of bias, variance, sensitivity and specificity, data for which we know the "truth" is required. Statistical challenges include taking into account effects of background noise and appropriate normalization of the data. For GeneChip data, some different models to normalize signal values or normalize probe pair values have been proposed (Bolstad et al., 2003; Geller et al., 2003; Irizarry et al., 2003a; Irizarry et al., 2003b; Li and Hung Wong, 2001; Stuart et al., 2001).

In order to be combined across studies, quantitative estimation must address the same measure or quantity, be standardized to the same scale, and include some measure of variability (Geller et al., 2003; Knudtson et al., 2006; Tusher et al., 2001). Moreover with the increasing awareness and usage of GeneChip technology and willingness to continue to use GeneChip software among many biologists, it is worth improving the performance or correcting the problems of the software.

The normalization method used in GeneChip software is called scaling and is defined as an adjustment of the average signal value of all arrays to a common value, the target signal value, in order to make the data from multiple arrays comparable (Affymetrix, 2002; http://www.affymetrix.com/index.affx). Bolstad and Jonssson reviewed these methods and find *quantile normalization* to perform best (Bolstad and Jonsson, 2002). The goal of quantile normalization is to make the distribution of probe intensities the same for arrays. The normalization maps

probe level data from all arrays $i = 1, \ldots, I$ so that an I-dimensional quantile-quantile plot follows the I dimensional identity line. The risk to remove some of the signal in the tails could be a problem of this approach. However, empirical evidences suggest this is not a problem in practice (Bolstad et al., 2003). In general, algorithms that affect statistical analysis include:

1. *Image analysis*: gridding, spot recognition of the scanned image (segmentation algorithm), removal or marking of poor-quality and low-intensity features (called flagging).
2. *Data processing*: background subtraction (based on global or local background), determination of spot intensities and intensity ratios, visualization of data (i.e. see MA plot), and log-transformation of ratios, global or local normalization of intensity ratios.

Statistical analysis: high-level investigation

After removing the bad quality data, it is the time of exploration of reliable data. Because a typical microarray experiment contains a large number of hypotheses and a limited number of replicates, high false-positive rates are a common problem in the identification of *Differential Expressing Genes (DEGs)*. An important factor in minimizing false positives is to incorporate an appropriate error model into the signal/noise metric. Therefore, the good quality data are further filtered so that only the genes that show some changes in the expression during the experiment are preserved in the dataset. There is a number of novel and very complex tests that are available or are being developed for analysis of large data sets, such as microarray data, that are sufficiently robust to accurately determine statistical significance.

Most common statistical methods can be divided into one of these two categories: parametric statistics, which uses the numerical data, such as the arithmetic mean, standard deviation and other parameters to determine significant differences between sets of data; and non-parametric approaches,

such as Mann-Whitney U test (two groups) or Kruskal-Wallis test (two or more groups), that use ranks of numerical data rather than the data themselves. In the parametric tests assumptions regarding the normal distribution of the data and the quality of variance among the groups must be made. These assumptions are often sufficiently satisfied to make parametric statistics extremely useful and a feasible starting point for analysis. However, if data are generated from populations that do not meet these assumptions, these methods become unreliable because the mean and variance will no longer completely describe the population, therefore skewing of the data from non-normal variances will lead to false conclusions regarding the data set. Some series of high-level analysis that are commonly used are: T-test, ANOVA, Bayesian method or Mann-Whitney test, PCA, Clustering and other statistical methods that take the underlying structure of gene networks into account, representing either associative or causative interactions or dependencies among gene products (Dehmer et al., 2008; Dehmer and Emmert-Streib, 2008).

2.4 Advantages and Disadvantages

One of the biggest advantages of DNA microarray technology is that it can evaluate simultaneously the relative expression of thousands of genes by using small amounts of materials, providing gene signatures for particular physiological or pathological situations. In addition, the procedures can be easily automated. Furthermore, the capacity of measurement of gene expression by DNA microarray is huge, allowing researchers to take the expression of all genes from an individual into consideration i.e., for disease analysis in so called “personalized medicine”. One of the major disadvantages of DNA microarray technology is that it only evaluates gene expression at a transcriptional level, but often in the regulation of protein functions, other mechanisms as posttranscriptional modifications (i.e., phosphorylation) are involved. Moreover, a “hot” list is often provided by

statisticians to researchers, to describe which genes are mostly upregulated or downregulated, and those genes with minor or no changes in mRNA expression are often considered as not involved or not important by the researchers. However, there are numerous examples showing a disassociation between the abundance of mRNA and the level of translated protein for a gene of interest, and between the abundance of mRNA and the effect of the gene of interest on a particular biological process.

3. Methodologies

3.1 Data pre-processing and normalization

Microarrays measure the target quantity indirectly by measuring another physical quantity, the intensity of the fluorescence of the spots on the array for each fluorescent dye (see Figure 1). These images should be later transformed into the gene expression matrix (*image processing*). This task is not a trivial one because:

1. the spots corresponding to genes should be identified;
2. the boundaries of the spots should be determined;
3. the fluorescence intensity should be determined depending on the background intensity.

Following image processing, the data generated for the arrayed genes must be normalized before the identification of interesting genes. This process is necessary to adjust the variability that depends on the nature of the experimental design. In fact, different non-biological sources of variability must be identified and mitigated, before to consider the sources of biological variability that we need to estimate. In particular, in the GeneChip experimental process, the sources of variability are induced by the biological nature of experimental (interesting genes), the sample preparation (total RNA isolation, as well as labeling) and the system (instruments and arrays). As a result of the standardization of the hybridization, staining, washing, and scanning, as well as the quality controls built into manufacturing processes, system noise is not a significant source of technical variation and does not need to be addressed (Sherlock, 2001). However, without careful technique and planning, sample preparation can be a large, unexpected, and unnecessary source of variation. The objective of normalization is to adjust the gene expression values of all genes on the array so that the genes that are not really differentially expressed have similar values across the arrays.

Data pre-processing steps, which combine multiple probe signals into a single absolute call, are known as *normalization procedures*. They usually involve three steps: (a) background adjustment, (b) normalization and (c) summarization (Gautier et al., 2004). The background adjustment is defined as the process of correcting probe intensities on an array using information only on that array. The normalization steps is the process of removing non biological variability between arrays and the summarization is the process of combining the preprocessed PM probes together to compute an expression measure for each probe set on the array. Different methods have been devised for each of the three steps and thus a great number of possible combinations exist, facing the microarray user community with a complex and often daunting set of choices.

The performance of a normalization method would then be ranked based on the overall error estimate in the prediction of the concentration of these mRNAs (Bolstad et al., 2003; Liu et al., 2005). Among the normalization methods used for the common Affymetrix GeneChip (see Figure 3), the *Robust Multiarray Average (RMA)* method and the statistical algorithm implemented in *Affymetrix's Microarray Suite (MAS5)* program are considered the gold standard to control for systematic variation in samples of unrelated individuals (Bolstad et al., 2004; Chesler et al., 2005; Irizarry et al., 2003b).

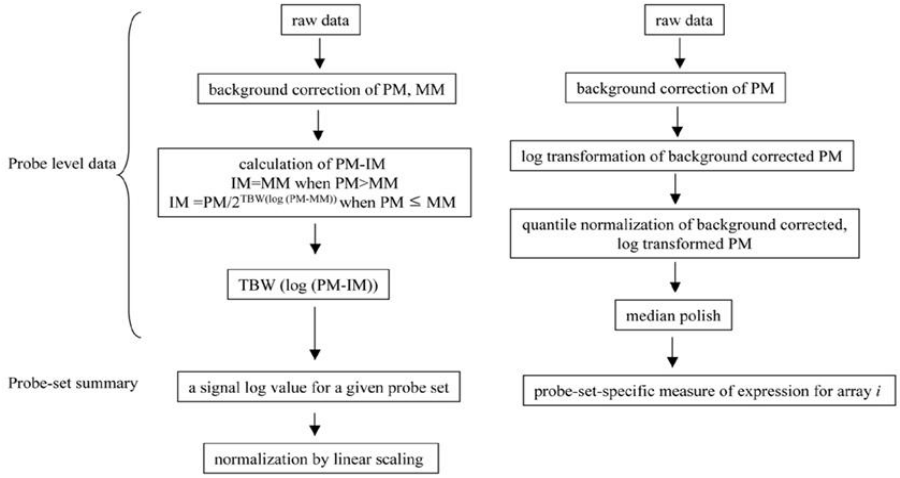

Figure 3: ***Normalization strategies for Affymetrix GeneChip data. A)*** *MAS5 normalizes the value of probe-set summary by linear scaling based on a reference array.* ***B)*** *RMA (robust multi-array average) normalizes the value of each probe by quantile normalization in multiple arrays (Do and Choi, 2006).*

3.1.1 MAS5 method

Affymetrix Microarray Suite 5.0 (MAS5) is a program created by Affymetrix and it determines gene expression intensity by applying a Tukey's Biweight algorithm (Hoaglin DC, 1983) to determine probe set intensity. This algorithm combines signals from the multiple PM and MM probes that target each transcript into a single value that sensitively and accurately represents its concentration, based on the p-value. MAS5 does this by calculating a robust average of the (logged) PM-MM values (Hubbell et al., 2002), increased variation is observed at low signal strengths and is at least in part due to the extra noise generated by subtracting the MM values from their PM partners (Irizarry et al., 2003b). MAS5 normalizes by picking specific regions within the GeneChip and adjusting the signal intensities for each probe to a user defined value.

Background correction

MAS5 performs background correction using neighboring probe sets. The entire array area is divided into 16 rectangular zones and the lowest 2nd percentile of the probe values are chosen to represent the background value in given zones (Draghici et al., 2003). Then, the background value is computed as a weighted sum of the background values of the neighboring zones with the weight being inversely proportional to the square of the distance to a given zone. The negative value by subtraction of the position specific background is avoided with a small threshold value.

Summarization, Normalization and Detection of Call

MAS5 returns two values, the first, an estimate of transcript concentration, and the second, a measure of how much the software "believes" the first. This value is referred to as the *detection p-value* and it is subsequently used to generate a *detection call*, which flags the transcript as "Present", "Marginal" or "Absent" (P/M/A). Probe intensities for each probe set should be summarized to define a measure of expression representing the amount of the corresponding mRNA species. Specifically the MAS5 signal (measure) is defined as

$$\text{signal} = \text{Tukey Biweight } [\log(\text{PM}_{ij} - \text{CT}_{ij})] \qquad \textbf{(Eqn.1)}$$

with CT_{ij} a quantity derived from MM, that is never bigger than its PM pair (represent Change Threshold), on array *i*. Each of these measures relies upon the difference $PM - MM$ with the intention of correcting for nonspecific binding. In more detail it is defined a error term as

$$\log(\text{PM}_{ij} - \text{CT}_{ij}) = \log \theta_i + \varepsilon_{ij} \qquad \textbf{(Eqn.2)}$$

where ε_{ij} and θ_i, represent random error term and expression quantity, respectively. It is calculated as the anti-log of a robust average (Tukey biweight) of the values of log (PM_{ij} - CT_{ij}). To avoid taking the log negative numbers, *CT* is defined as a quantity equal to MM when MM < PM, but *CT*

is replaced with *PM x Tb (MM/PM)*, where *Tb* is the function of Tukey's bi-weight. See (Hubbell, 2001) for more details.

MAS5 uses, for the normalization among multi-array experimental datasets, a simple linear scaling not on the probe level intensity but on the summarized gene-level intensity. This approach is not effective on the dataset whose probe level intensity distribution contains large chip by chip differences (Zhang, 2004). This software gives the user the ability either to choose a particular target signal value or to choose a particular baseline array against which to normalize. Based on this choice, the normalization signal log values (SLV) are calculated as follows:

$$SLV_{ij} - \log_2(sf_i) \qquad \textbf{(Eqn.3)}$$

where SLV denotes the signal log value for gene *j* on array *i* given by Eqn. 2 and sf_i is the scaling factor for array *i*. The signal value reported by the MAS5 software is 2 raised to the power of normalized signal log value. Affymetrix software computes the scaling factor as

$$sf_i = \log_2(Sc) - \log_2\left(\text{TrimMean}_j(2^{SLV_{ij}})\right) \qquad \textbf{(Eqn.4)}$$

where *Sc* is the target constant used for normalizing all arrays. The "TrimMean" notation denotes the mean of the signal values on array *i* for the housekeeping genes, excluding outliers in the upper and lower 2% of the distribution. Affymetrix computes the trimmed mean on the absolute intensity scale. Because SLV is a $\log_2$ transformed value, the value is raised to the power 2 before taking the trimmed mean. If no housekeeping genes are identified, then the trimmed mean uses the signals for all of the genes on the array.

In addition to expression summaries, the *Call* information is also generated. Detection calls are used to determine whether the transcript of a gene is detected (present) or undetected (absent), and it is make up by the intensity difference of PM and MM probe cells. The original approach to data analysis,

proposed by the manufacturer, was to use the MAS5 expression summary to provide an estimate of transcript concentration, alongside detection calls to filter out unreliable probe sets. Despite the fact that the expression summary algorithm has been shown to perform poorly on the test datasets described above, many researchers have continued to use this combined strategy to process their data. Several different methods have been used to make detection calls (Liu et al., 2002; Lockhart et al., 1996). The most widely used and robust results to use a signed rank test to consider the significance of the difference between the PM and MM values for each probe set (Liu et al., 2002).

3.1.2 RMA method

Robust Multichip Average (RMA) is one of the most popular algorithms for pre-processing probe level data from oligonucleotide arrays. In the RMA model, it is assumed that the observed PM intensities are the sum of noise (considered to be normally distributed) and signal (exponentially distributed), and the information of the MM intensities is ignored, which cause more variance (Cope et al., 2004). RMA provides a greater than five-fold reduction of the within-replicate variance as compared to other methods, provides more consistent estimates of fold change, and provides higher specificity and sensitivity when using fold change analysis to detect DEGs (Irizarry et al., 2003a). RMA is unique in that it adjusts for background noise, performs a quantile normalization, transforms data into a $\log_2$ and then summarizes the multiple probe into one intensity (Bolstad et al., 2004; Bolstad et al., 2003; Cope et al., 2004; Irizarry et al., 2003a; Irizarry et al., 2003b).

Background correction

Irizarry (Irizarry et al., 2003b) conducted a global background correction by signal and noise (background) convolution model. The background is assumed to be additive, so that the intensity of PM probe is a sum of background and foreground (spot) intensities. In particular, PM intensity distribution is modeled by an exponentially distributed signal component S with pa-

rameter λ, and a normally distributed background component B with mean μ and standard deviation σ. E(S|PM) represents background corrected value of each PM. φ and Φ are the normal density and cumulative density, respectively. Positive signal components are estimated after adjustment of the background components. It is also assumed that error in intensity values is multiplicative, i.e., the larger the absolute intensity value, the larger the error.

$$PM = S + B \quad \textbf{(Eqn.5)}$$

$$S \sim exp(\lambda) \quad \textbf{(Eqn.6)}$$

$$B \sim N(\mu, \sigma) \quad \textbf{(Eqn.7)}$$

$$E(S|PM) = PM - \mu - \lambda\sigma^2 \quad \textbf{(Eqn.8)}$$

$$\sigma = \frac{\phi\left(\frac{PM-\mu-\lambda\sigma^2}{\sigma}\right)-\phi\left(\frac{\mu+\lambda\sigma^2}{\sigma}\right)}{\Phi\left(\frac{PM-\mu-\lambda\sigma^2}{\sigma}\right)-\Phi\left(\frac{\mu+\lambda\sigma^2}{\sigma}\right)} \quad \textbf{(Eqn.9)}$$

Normalization

After background correction, the normalization of GeneChip data can be applied onto probe levels as well as onto gene expression measures depending on normalization strategies. RMA adopted probe level quantile normalization that makes the distribution of probe intensities for each array in a set of arrays by taking the mean quantile and substituting it as the value of the data item in the original dataset. The goal of the quantile method is to make the same distribution of probe intensities for each array in a set of arrays. A quantile-quantile plot shows that the distribution of two data vectors is the same if the plot is a straight diagonal line and not the same if it is other than a diagonal line. Also the quantile normalization has been shown to have the best performance and works by making the distribution of intensities at the probe level (Bolstad et al., 2003). The *quantile normalization* algorithm appears to more performance and also is less noisy than all other methods (Bolstad and Jonsson, 2002; Irizarry et al., 2003b). Quantile normalization

method assumes that the distribution of intensity values is similar on every chip.

One possible problem of this method is that it forces the quantile values to be equal. This would be most problematic in the tails where it is possible that a probe could have the same value across all the arrays. However, since probe set expression measures are typically computed using the value of multiple probes (Bolstad et al., 2003).

Summarization of probe set

Expression values are calculated from rearranged mean (normalized) values using median polishing. Median polishing is an iterative method, which aims to centralize both column medians and row medians to one. The summarization used is motivated by the assumption that observed log-transformed PM values follow a linear additive model containing a probe affinity effect, a gene specific effect and an error term. In the case of RMA, median polishing works on a matrix were every row corresponds to one gene and every column to one chip. The median for every row is calculated and the median is subtracted from the intensity values so that the row median becomes one. Next, the median for every column is calculated and the median is subtracted from the intensity values so that the column median becomes one. RMA implementations work with log-transformed data, so at the median polishing phase the row and column medians are actually centered to zero. A common early step in microarray data analysis is log transformation. Log transformation has several important effects: a) allows normalization of data even though error in Signal intensity increases as the magnitude of Signal intensity increases; b) makes data more symmetrical; and c) reduces the influence of a single measurement (Durbin et al., 2002).

3.2 Analysis of Variance (ANOVA)

Microarray data can be interrogated using *ANalysis Of VAriance (ANOVA)*, a powerful and general method of data analysis that has been extensively developed and studied for more than 75 years (Fisher, 1925). Analysis of Variance is one of the most commonly used multivariate statistic method, and its purpose is to test for significant differences between the means of several groups (Gelman, 2005). The ANOVA will give the same result as the t-test, when two means are compared. However, unlike the t-test, ANOVA does not specify which of the groups are significantly different from each other, but it only determines that there are significant differences. In the ANOVA, the effect size is the difference between the two populations divided by estimated population standard deviations. Subsequently a p-value is generated and can be used to determine the significance of results.

3.2.1 Problem definition and model assumptions

Let us consider an experiment measuring the expression level of a given gene in a number of *k* conditions. Each gene *i* is measured *n* times for a total of measurements of

$$\sum_{i=1}^{k} n_i \qquad \textbf{(Eqn.10)}$$

The basic question is to decide whether there is any difference in the expression level of the given gene between the *k* conditions. Under the null hypothesis that the different conditions are not really different and, therefore, all measurements actually come from a single distribution. In these conditions, all means would be the same:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \qquad \textbf{(Eqn.11)}$$

The alternative hypothesis is that there are at least two means that are different from each other. This particular data layout and set of hypotheses is characteristic to a Model I (or fixed effects) ANOVA. In this model, the specific interest is the differences between any pair of the specific conditions

considered. If such differences exist, it is of interest to identify which are the elements that differ from each other. Furthermore, in this data layout, each measurement belongs to single group. In other words, the data was factored one-way. This analysis is called a one-way ANOVA and it is a parametric test, that involve a number of assumptions as follows:

1. The *k* samples are independent random samples drawn from *k* specific populations with means $\mu_1, \mu_2, \ldots, \mu_k$;
2. All *k* populations have the same variance σ^2 (homoscedasticity);
3. All *k* populations are normal.

The strictest assumption, for which the test is not robust, is the assumption of independence (Boneau, 1960). Another main problem is that the statistical hypothesis testing does not provide absolute conclusions. Instead, each time a null hypothesis is rejected, there is a non zero probability of the null hypothesis being actually true. This is the probability of a Type I error (or significance level). When many of such tests are carried out for purpose of drawing a single conclusion, a single mistake in each one of the individual test is sufficient to invalidate the conclusion. Thus, the probability of a Type I error increases with the number of tests even if the probability of Type I error in each test is bounded by the chosen level of significance.

Briefly summarizing, ANOVA is a statistical tool used to identify differences between experimental group means. ANOVA is commonly used in experimental designs with one dependent variable that is a continuous numerical parametric outcome measure and multiple experimental groups within one or more independent (categorical) variables. These independent variables are called factors and groups within each factor are also referred to as levels. Depending on the number of factors included in the model, it is possible distinguish one way and *n*-way ANOVA. One-way ANOVA can be

used when the researcher wants to examine the influence of only one independent variable (factor) on the dependent variable. At its simplest, a one-way ANOVA can be used to test the hypothesis that some variable of interest differs among groups (one factor); two-way ANOVA can test for differences among groups while controlling for other categorical variables (two factors); and thus extending (*n*-factors).

3.2.2 One-way ANOVA

The null hypothesis tested by one-way ANOVA is that two or more population means are equal. The question is whether (H_0) the population means may equal for all groups and that the observed differences in sample means are due to random sampling variation, or (H_n) the observed differences between sample means are due to actual differences in the population means. The general idea behind one-way ANOVA in microarray experiment is every measurement in a microarray experiment is associated with a particular combination of an array in the experiment, a variety, and a gene. The measurements of each array (condition) vary around their mean. This is a variability within group and will be characterized by a corresponding within group variance. At the same time, the means of each treatment will vary around an overall mean. This is due to an inter-group variability. Finally, as result of two above, each individual measurement varies around the overall mean. These mean that ANOVA is to study the relationship between the inter-group and within-group variabilities (or variances) (Nickerson, 2000).

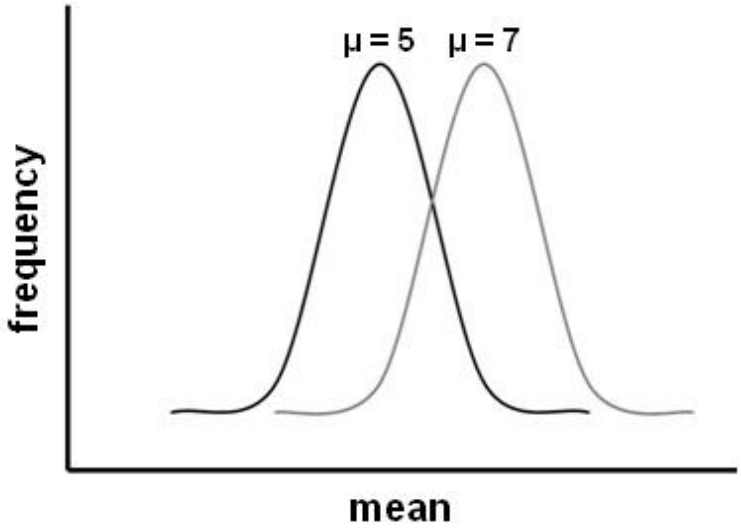

Figure 4: ***Representative example of ANOVA.*** *It tests if the central tendency (mean) is different between samples. In the figure, there are two subpopulations representing two hypothetical samples with the normality assumption (normal bell-shaped curve). The y-axis measures the variable (given by frequency) and the x-axis the mean of samples.*

Figure 4 shows two samples with means μ as 5 and 7. In this case, the variability within groups is much smaller than the overall variability. Thus this may allow to reject the hypothesis that the two samples were drawn from the same distribution and so to accept that there is a significant difference between the two samples. In general, the ANOVA method seeks to detect sources of variation in the values of dependent variable and divide the total variability into components associated with each source. The total variability is the sum of squared deviations of each measurement from the overall mean and can be decomposed into a sum of squares (SS) due to suspected sources of variation (model sum of squares) and a sum of squares (SS) resulting from the error:

$$SS\ (Total) = SS\ (Model) + SS\ (Error) \qquad \textbf{(Eqn.12)}$$

The first step is to check the data to make sure that the raw data are correctly assembled and that assumptions have not been violated in a way

that makes the test inappropriate. From the assumption 2, the variance is the same within the two populations. An unbiased estimate of this common population variance can be calculated separately from each sample. The numerator of the variance formula is the sum of squared deviations around the sample mean, or simply the sum of squares for sample *j* (abbreviated as SS_j). The denominator is the degrees of freedom for the population variance estimate from sample *j* (abbreviated as df_j).

$$\textit{Unbiased estimate of } \sigma_j{}^2 = \frac{\sum_i (y_{ij} - \bar{y}_j)^2}{(n_j - 1)} = \frac{SS_j}{df_j} = s_j{}^2 \qquad \textbf{(Eqn.13)}$$

To pool two or more sample estimates of a single population variance, each sample variance is weighted by its degrees of freedom. This is equivalent to adding together the sums of squares for the separate estimates, and dividing by the sum of the degrees of freedom for the separate estimates.

$$\textit{Pooled estimate of } \sigma_y{}^2 = \frac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}{(n_1 + n_2 - 2)} = \frac{SS_1 + SS_2}{df_1 + df_2} = s_y{}^2$$

(Eqn.14)

Subsequently, it is necessary to examine a second approach to testing two means for equality. The logic of this approach extends directly to one-way analysis of variance with k groups. These two estimates are expected to be equal if the population means are equal for all k groups (Equation 11), but the estimates are expected to differ if the population means are not all the same. In ANOVA terminology, the numerator of Equation 14 is called the *Sum of Squares Within Groups* (SS_{WG}) and the denominator is called the *degrees of freedom Within Groups* (df_{WG}). The estimate of the population variance from Equation 14, SS_{WG}/df_{WG}, is called the *Mean Square Within Groups* (MS_{WG}). Equation 15 is an equivalent way to express and compute MS_{WG} as

$$\textit{Within-groups estimate of } \sigma_y{}^2 = \frac{\sum_{ij}\left(y_{ij}-\bar{y}_j\right)^2}{\sum_j\left(n_j-1\right)} = \frac{SS_{WG}}{df_{WG}} = MS_{WG}$$

(Eqn.15)

If the null hypothesis is true and the assumptions are valid (random, independent sampling from normally distributed populations with equal variances), then a second independent estimate of the population variance can be calculated. As is stated by the *Central Limit Theorem*, if independent samples of size *n* are drawn from some population with variance, an unbiased estimate of the variance for the distribution of all possible sample means (for samples of size *n*) are calculated as:

$$est\ \sigma_j{}^2 = s_{\bar{y}}^2 = \frac{\sum_j\left(\bar{y}_j-\bar{y}_{..}\right)^2}{(k-1)} \qquad \textbf{(Eqn.16)}$$

Calculation of this second estimate of the population variance using ANOVA notation is shown in Equation 17. The MS_{BG} is the best estimate of the population variance based only on knowledge of the variance among the sample means. Equation 17 allows for unequal sample sizes and it is computed as

$$\textit{Between-groups estimate of } \sigma_y{}^2 = \frac{\sum_j n_j\left(\bar{y}_j-\bar{y}_{..}\right)^2}{(k-1)} = \frac{SS_{BG}}{df_{BG}} = MS_{BG}$$

(Eqn.17)

The ANOVA is performed using F statistic based on the ratio of between-group to within-group variance estimates:

$$F = MS(Model) / MS(Error) \qquad \textbf{(Eqn.18)}$$

The F ratio is designed as

$$F(df_{BG}, df_{WG}) = \frac{Between-groups\ estimate\ \text{of}\ {\sigma_y}^2}{Within-groups\ estimate\ \text{of}\ {\sigma_y}^2} = \frac{MS_{WG}}{MS_{BG}}$$

(Eqn.19)

When more than two group means are compared in ANOVA (i.e. there are three treatment groups in the study), the F statistic will only tell us whether there are significant differences in the group means as a whole. It will not tell us what are the differences between each groups and which group means differ from each other. Thus, ANOVA is usually followed up with a multiple comparison procedures with the purpose of identifying which group means differ from each other.

3.2.3 Two-way ANOVA

In a one-way ANOVA, the effects of various levels or treatment conditions of one independent variable on a dependent variable are examined. Many experimental designs can be established to test the effect that two variables may have on a data set. For example, it may examine normal *vs.* tumor cells, along with the effect of two different drugs, making a total of four different sample sets. In this case a two-way ANOVA can be used to investigate differences in gene expression between the different conditions, as well as type of cells differences within and between each condition. In this case, two separate ANOVAs cannot adequately examine the possible interactions that can be generated between the two variables, and, so, a two-way analysis of variance is the best methodology. A two-way ANOVA consists of three significance tests: a test of each of the two main effects and a test of the interaction of the variables.

3.3 Principal Component Analysis (PCA)

In gene expression experiments each gene and each experiment may be represented as dimension. For example, a set of 10 experiments involving 20000 genes may be conceptualized as 20000 data points (gene) in a space with 10 dimensions (experiments) or also 10 points (experiments) in a space with 20000 dimensions (genes). This simple example gives a clear visualization of the problem of the large number of dimensions that lies into the nature of microarray experiments. Different approaches try to reduce the number of dimensions and also the complexity of problem. A common statistical approach is to pay attention to those dimensions that account for a large variance in the data and to ignore the dimensions in which the data do not vary much. This is the approach used by Principal Component Analysis (PCA) (Ringner, 2008).

PCA is an exploratory multivariate statistical technique for simplifying complex data sets (Basilevsky, 1994; Hoaglin et al., 1983). Given *m* observations on *n* variables, the goal of PCA is to reduce the dimensionality of the data matrix by finding *r* new variables, where *r* is less than *n*. Termed principal components, these *r* new variables together account for as much of the variance in the original *n* variables as possible while remaining mutually uncorrelated and orthogonal. Each principal component is a linear combination of the original variables, and so it is often possible to ascribe meaning to what the components represent. Therefore, the principal components are linear combinations of random or statistical variables which have special properties in terms of variance. Principal components analysis has been used in a wide range of biomedical problems, including the analysis of microarray data in the search of outlier genes (Hilsenbeck et al., 1999) as well as the analysis of other types of expression data (Craig et al., 1997; Vohradsky et al., 1997) and also in the cluster analysis of data (Peterson, 2003). A PCA analysis of DNA microarray data can consider either the genes or the expe-

riments as variables. When genes are variables, the analysis creates a set of "principal gene components" that indicate the features of genes that best explain the experimental responses they produce. When experiments are the variables, the analysis creates a set of "principal experiment components" that indicate the features of the experimental conditions that best explain the gene behaviors they elicit. When both experiments and genes are analyzed together, there is a combination of these effects, the utility of which remains to be explored. By using a few components, each sample can be represented by relatively few numbers instead of by values for thousands of variables. Samples can then be plotted, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped.

Descending in more detail, PCA calculates a new system of coordinates (principal components). To compute they, the *n* eigenvalues and their corresponding eigenvectors are calculated from the *n*x*n* covariance matrix of conditions. Each eigenvector defines a principal component (PC). A component can be viewed as a weighted sum of the conditions, where the coefficients of the eigenvectors are the weights. The projection of gene *i* along the axis defined by the *jth* principal component is:

$$a'_{ij} = \sum_{t=1}^{n} a_{it}\, v_{tj} \qquad \textbf{(Eqn.20)}$$

Where v_{tj} is the *tth* coefficient for the *jth* principal component; a_{it} is the expression measurement for gene *i* under the *tth* condition. *A'* is the data in terms of principal components. Since *V* is an ortho normal matrix, *A'* is a rotation of the data from the original space of observations to a new space with principal component axes. The variance accounted for by each of the components is its associated eigenvalue; it is the variance of a component over all genes. Consequently, the eigenvectors with large eigenvalues are the ones that contain most of the information; eigenvectors with small eigenvalues are uninformative.

In intuitive terms, the covariance matrix captures the shape of the set of data points. In Figure 5A is illustrated an *n*-dimensional hyper-ellipsoid which includes the data. The eigenvectors of covariance matrix, or the directions found by the PCA, will be the directions of the main axes of the ellipse and most of the variability in the data lies along a one-dimensional space that is described by the first principal component (PC1). In this example the second principle component (PC2) can be discarded because the first principle component captures most of the variance present in the data. The essential aspect of the PCA is related to the fact that the absolute value of eigenvalues are directly proportional to the dimension of the multidimensional ellipse in the direction of corresponding eigenvector. Deciding how many and which components to use in the subsequent analysis is a major challenge that can be addressed in several ways (Khan et al., 2001; Landgrebe et al., 2002; Saal et al., 2007). In general, as shown in Figure 5B the first principal components are able to capture the most of the variance of the data.

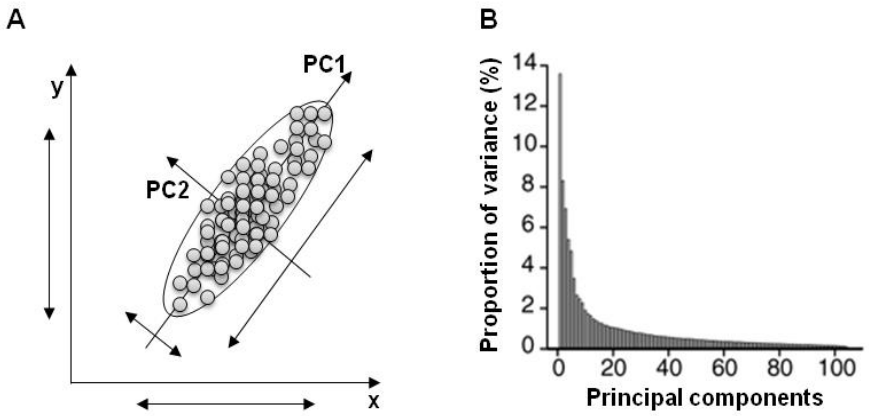

Figure 5: ***Example of PCA analysis. A)*** *In this data set most of the variance is along the first Eigenvector (PC1) with a small variance along the second direction (PC2) being probability due to the noise. The PCA will find a new coordinate system in which the first coordinate is the direction on which the data have maximum variance (the first eigenvector), the second coordinate is perpendicular on the first and captures the second largest variance, etc.* ***B)*** *The variance of the principal components.*

The first application of PCA is to explore high-dimensional data sets, as outlined above. Most often, three-dimensional visualizations are used for such explorations, and samples are either projected onto the components, as in the examples here, or plotted according to their correlation with the components (Alter et al., 2000). The principal components are uncorrelated and they may represent different aspects of the samples. This suggests that PCA can serve as a useful first step before clustering or classification of samples, and more in general PCA can potentially provide insights into different choices of pre-processing and variable selection.

In spite of its usefulness, PCA has also drastic limitations. Those limitations are mainly related to the fact that PCA only takes into consideration the variance of the data that is a first order statistical characteristic of the data. Another major limitation is that PCA takes into account only the variance of the data and completely discards the class of each data point. In some cases, such handling of the data will not produce the required results as the classes would not be defined by the PCA. Furthermore, PCA may fail to distinguish between classes when the class variance is the same. PCA's limitations may be overcome by alternative approaches by using higher order statistical dependencies as Skew and Kurtosis (Hyvarinen and Oja, 2000).

3.4 Clustering Algorithms

Clustering is a classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult combinatorial problem, and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur. Cluster analysis is currently the most frequently used multivariate technique to analyze gene sequence expression data. Clustering has become so popular in

this field, (Ben-Dor et al., 1999; Claverie, 1999; Zhang, 1999) and it is so great that sometimes clustering is mistakenly taken as a very fuzzy and all-inclusive ultimate goal of microarray data analysis. In fact, clustering is the process of grouping together similar entities and it can be done on any data: genes, samples, time points in a time series, etc. This kind of statistical inference is particularly important in the context of analyzing high-dimensional genomic data sets. The particular type of input makes no difference to the clustering algorithm, in fact all inputs are a set of n numbers or an n-dimensional vector. It is obviously unquestionable the great strength of this approach, that has the ability, for example, to identify without a priori knowledge about the data gene sets that show similar patterns of expression. Therefore it is possible to exploit this potential only after making several designing choices carefully. Before comparing the clustering methods, it is important to start to define the meaning of similarity.

3.4.1 Measure of similarity

In the mathematical language a measure of similarity is called *distance* or *metric*. A distance is a formula that takes two points in the input space of the problem and calculates a positive number that contains information about how close the two points are to each other. The input space of the problem is a n-dimensional space so the two points can be for instance two measured across n experiments, each represented by the expression values of n genes. In other words, we define an "expression vector" for each gene that represents its location in the "expression space". Any function d that satisfies the following tree properties is termed a distance:

1. no-negativity $d(x,y) \geq 0$;
2. symmetry $d(x,y) = d(y,x)$;
3. identification mark $d(x,x) = 0$.

There are many different ways in which such a measure of similarity can be calculated. Because there are many different types of data (i.e., ordinal, nominal, continuous) and approaches for analyzing these data, the literature on distances is quite broad. References that considerer the application of distances in either clustering or classification include: (Kaufman and Rousseeuw, 1990) and (Duda et al., 2001). In the following paragraphs the common principal distances used in the analysis of gene expression data are introduced (Johnson, 1998).

Euclidean distance

The Euclidean distance (De Smet et al., 2002; Wang et al., 2002) between two *n*-dimensional vectors, $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$, is:

$$d_E(\mathrm{x},\mathrm{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \textbf{(Eqn.21)}$$

This is the useful distance that is used for most of the practical purposes. Its numerical value comes from Pythagorean Theorem, where x_i and y_i are the measured expression values, respectively for gene X and Y in the experiment *i*, and the summation runs over the *n*-experiments under analysis. It works well when a data set has "compact" or "isolated" clusters (Mao and Jain, 1996).

Standardized Euclidean distance

All the distances so far considered give exactly the same importance to all dimensions. The idea behind standardized Euclidean distance is that not all directions are necessarily the same relevance. The standardized Euclidean distance takes this into consideration by dividing the distance of each dimension by the standard deviation of each dimension. The standardized Euclidean distance between two *n*-dimensional vectors, $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$, is:

$$d_E(\mathrm{x},\mathrm{y}) = \sqrt{\frac{1}{s_1^2}(x_1 - y_1)^2 + \cdots + \frac{1}{s_n^2}(x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} \frac{1}{s_1^2}(x_i - y_i)^2}$$

(Eqn.22)

Manhattan distance

The Manhattan distance between two *n*-dimensional vectors, $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$, is:

$$d_E(\mathrm{x},\mathrm{y}) = |x_1 - y_2| + |x_1 - y_2| + \cdots + |x_n - y_n| = \sum_{i=1}^{n} |x_i - y_i|$$

(Eqn.23)

where $|x_i - y_i|$ represents the absolute value of the difference between x_i and y_i . The Manhattan distance (also called city-block) is the distance that one needs to travel in an environment in which one can move only along directions parallel to the x and y axes, in other words no diagonal movements are possible. The Manhattan distance is independent on the path travelled between the two points.

Chebychev distance

The Chebychev distance between two *n*-dimensional vectors, $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$, is:

$$d_{max}(\mathrm{x},\mathrm{y}) = \max_i |x_i - y_i| \qquad \textbf{(Eqn.24)}$$

The Chebychev distance simply picks the largest difference between any two corresponding coordinates. For instance if the vectors $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ are two genes measured in *n*-experiments each, the Chebychev distance will pick the experiment in which these two genes are more different and will consider the distance between them as the effective value of distance between the genes. It is used when the goal is to reflect any big difference between any corresponding coordinates.

Correlation distance

The Pearson correlation distance (D'Haeseleer, 2005; Jiang, 2003; Yang, 2002) between two *n*-dimensional vectors, x = (x_1, x_2, ... , x_n) and y = (y_1, y_2, ... , y_n), is:

$$d_R(\mathrm{x},\mathrm{y}) = 1 - r_{xy} \qquad \textbf{(Eqn.25)}$$

where r_{xy} is the Pearson correlation coefficient of vectors x and y:

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x}\sqrt{s_y}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad \textbf{(Eqn.26)}$$

The Pearson correlation coefficient varies only between -1 and 1, so the distance $(1- r_{xy})$ will take values between 0 and 2. The Pearson correlation focuses on whether the coordinates of two points change in the same way. The magnitude of coordinates is less important since the denominator will be proportional to the magnitude of the vectors. If the vector is a set of measurements of given genes in a particular experiment and two such experiments are compared, the Pearson distance will be high if the genes vary in a similar way in the two experiments even if the change in the magnitude of the coordinates differs greatly. One drawback of Pearson's correlation coefficient is that it is not robust with respect to outliers (Heyer et al., 1999).

Distances are an integral part of all machine-learning algorithms and hence play a central role in the analysis of most experimental data. The distance that is used for any particular task can have a profound effect on the output, in fact the clustering procedure by a given algorithm is highly dependent on the distance metric used. Changing the distance metric may affect dramatically the number and membership of the clusters as well as the relationship between them. Both Euclidian distance and Pearson's correlation coefficient seem to work well as distance measure. In additional Euclidian distance may be more appropriate for log ratio data, while

Pearson's correlation coefficient may be better for absolute vaulted data (Gibbons and Roth, 2002).

3.4.2 Algorithms

Various clustering techniques have been applied to the identification of patterns in gene expression data, here are described only the methods used in this thesis. In general the clustering techniques can be described and classified in different ways (Jain and Dubes, 1988) for instance, as divisive or agglomerative. A divisive method begins with the clusterization of all elements in one cluster that is gradually broken down into smaller and smaller clusters. Agglomerative techniques start with single-member clusters which are gradually fused together. Further, clustering can be either supervised or unsupervised. Supervised methods use existing biological information about specific genes that are functionally related to "guide" the clustering algorithm. More in general supervised methods assign some predefined classes to a data set, whereas in unsupervised methods no prior assumptions are applied.

The principal clustering techniques used are Hierarchical clustering, *K*-means, Self-Organizing maps (SOMs) and Principal Component Analysis (PCA). This last will be only briefly described here, because it will be treated in more detail in a separate paragraph. Moreover, the existing literature is very rich in papers concerning alternative clustering methods and algorithms as well as their applications (Cho et al., 2001; Getz et al., 2000; Hastie et al., 2000; Herrero et al., 2001; Michaels et al., 1998; Yeung et al., 2001).

Hierarchical clustering

Starting from hierarchical clustering algorithms derive a nested series of partitions of data points. It has been used since the very beginning of microarray field (Eisen et al., 1998; Heyer et al., 1999) its major advantage is that it is

simple and the result can be easily visualized (Eisen and Brown, 1999). In fact the result of hierarchical clustering is a complete tree with individual patterns (gene or experiments) as leaves and the root as the convergence point of all branches which is also known as dendrogram. Example dendrograms are presented in Figure 6.

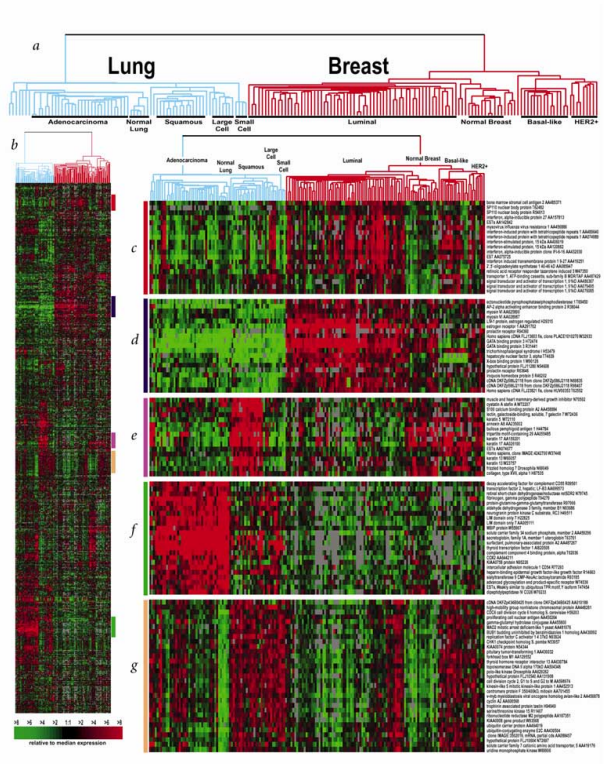

Figure 6: ***Hierarchical clustering analysis obtained by using the breast and lung carcinoma data sets from the publicly Stanford Microarray Database*** *(Chung et al., 2002).*

A dendrogram is a branching diagram representing a hierarchy of categories based on degree of similarity. In more detail, the hierarchical clustering is either an iteratively joining of the two closest clusters starting from single clusters (agglomerative, bottom-up approach) or an iteratively partition of clusters starting from the complete set (divisive, top-down approach), in which in the first step, the pair wise distance matrix is calculated for all of

the genes to be clustered. In the second step, the distance matrix is searched for the two most similar genes (or clusters) and each cluster consists of a single gene. This is the first true stage in the "clustering" process. If several pairs have the same separation distance, a predetermined rule is used to decide between alternatives. In the third step, the two selected clusters are merged to produce a new cluster that now contains at least two objects while in the fourth passage, the distances are calculated between this new cluster and all other clusters. There is no need to calculate all distances as only those involving the new cluster have changed. Last, steps 2-4 are repeated until all objects are in one bigger cluster. There are various hierarchical clustering algorithms that differ in the manner in which distances are calculated between the growing clusters and the remaining members of the data set (inter-cluster distance), including other clusters. The specification of distance between clusters is determined by the linkage method (Gibbons and Roth, 2002; Gordon, 1999):

1. *Single-linkage clustering*. The distance between two clusters is calculated as the minimum distance between a member of one cluster and a member of a second cluster. Consequently, it measures the distance between each member of one cluster to each member of the other cluster and takes the minimum of these. This technique produces trees with many long, single-addition branches representing clusters that have grown by aggregation.

2. *Complete-linkage clustering*. The distance between two clusters is calculated as the greatest distance between members of the relevant clusters (it calculates the distance between the furthest neighbors). This method tends to produce very compact clusters of elements and the clusters are often very similar in size.

3. *Centroid linkage*. Defines the distance between two clusters as the squared Euclidean distance between their centroids or means. This method tends to be more robust to outliers than other methods.

4. *Average-linkage clustering*. The distance between clusters is calculated using average values. This is calculated from the distance between each point in a cluster and all other points in another cluster. The two clusters with the lowest average distance are joined together to form a new cluster.

The selection of the linkage method to be used in the clustering greatly affects the complexity and performance of the clustering. Single or complete linkages require the less computations of the linkage methods. However, single linkage tends to produce stringy clusters that are not performance results. The centroid or average linkage methods produce better results regarding the accordance between the produced clusters and the structure present in the data. However both these methods require much more computations.

K-means

The *k* means algorithm is one of simplest and fastest partitional clustering method. As a consequence, it is widely used because of its simple implementation. It can be more effective than hierarchical methods (Tavazoie et al., 1999) if there is advanced knowledge about the number of clusters. However, it has a major drawback. The basic algorithm, as described by (MacQueen, 1967) begins with either an initial partition of the objects into *k* subgroups or an initial specification of *k* cluster centroids. As a result, the researcher has to assess the quality of the obtained clustering. In particular, in *k* means clustering, objects are partitioned into a fixed number *k* of clusters, such that the clusters are internally similar but externally dissimilar (see example in Figure 7).

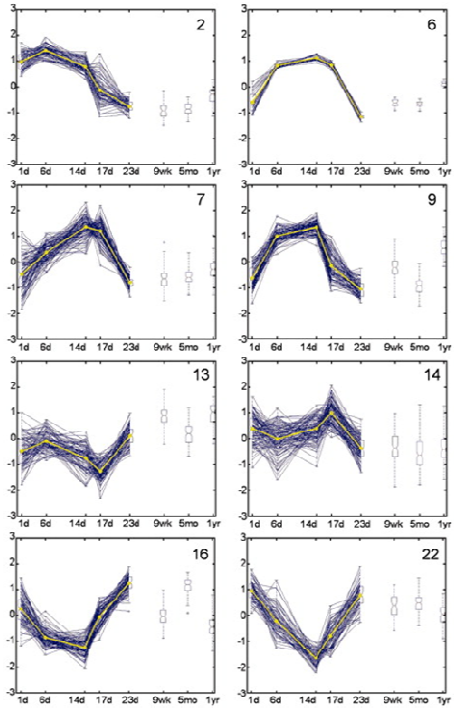

Figure 7: ***Cluster profiles obtained with K-mean cluster method****. For each x point, the standard deviation and the minimum and maximum values for each cluster are shown. Blue lines represent expression profiles for individual genes. Yellow lines are mean expression profile of clusters (Lin et al., 2004).*

The process involved is conceptually simple, but can be computationally intensive: first, all initial objects are randomly assigned to one of *k* clusters, it is also possible to estimate *k* from the data, taking the approach of a mixture density estimation problem; second, an average expression vector is then calculated for each cluster and this is used to compute the distances between clusters; third, using an iterative method, objects are moved between clusters and intra- and inter-cluster distances are measured with each move. Objects are allowed to remain in the new cluster only if they are closer to it than to their previous cluster; fourth, after each move, the expression vectors for each cluster are recalculated; and in the last step, the shuffling proceeds until moving any more objects would make the clusters more variable, increasing intra-cluster distances and decreasing inter-cluster

dissimilarity. More in general during the course of iterations, the program tries to minimize the sum, over all groups, of the squared within-group residuals, which are the distances of the objects to the respective group centroids. Convergence is reached when the objective function cannot be lowered any more. The obtained groups are geometrically as compact as possible around their respective centroids.

There are many way to assess the goodness of fit of a given clustering output, the principal is to compare the size of the clusters versus the distance to the nearest cluster. In this case if the inter-cluster distance is much larger than the size of the clusters, the cluster is deemed to trust worthier. Another possible quality indicator is the average of the distances between the members of a cluster and the cluster center or also the diameter of the smallest sphere including all members of a given cluster, but the last method can be more disadvantageous because the diameter of the smallest sphere including all members of the cluster is determined by the furthest pattern from the cluster. In consequence, this measure is sensitive to cluster outliers.

A major advantage of nonhierarchical clustering methods such as *k*-means is the computational feasibility. Unlike hierarchical methods, there is no need to compute or store all pairwise distances (or similarities) between objects. This makes possible to cluster a larger number of objects in less time. In the context of microarray data, this is particularly important for clustering genes, where they number may be of the tens of thousands. Many computer implementations of hierarchical clustering can not handle clustering of more than a few thousand genes. Disadvantages of *k*-means are that the method does require specification of a number of clusters and an initial partitioning, and the final results can be very sensitive to these choices.

Self-Organizing Maps (SOMs)

A self-organizing maps (SOMs), developed and studied by Kohonen is a neural net divisive clustering approach (Kohonen, 1995; Toronen et al., 1999) that uses unsupervised learning for which no prior knowledge of classes is required. A SOM assigns genes to a series of partitions on the basis of the similarity of their expression vectors to reference vectors that are defined for each partition. It is the process of defining these reference vectors that distinguishes SOMs from *k*-means clustering. SOMs are usually used to visualize and interpret large high-dimensional data sets. In SOM, every input is connected to every output via connections with variable weights. Also, the output nodes are highly interconnected. SOM tries to learn to map similar input vectors (gene expression profiles) to similar regions of the output array of nodes (Figure 8A).

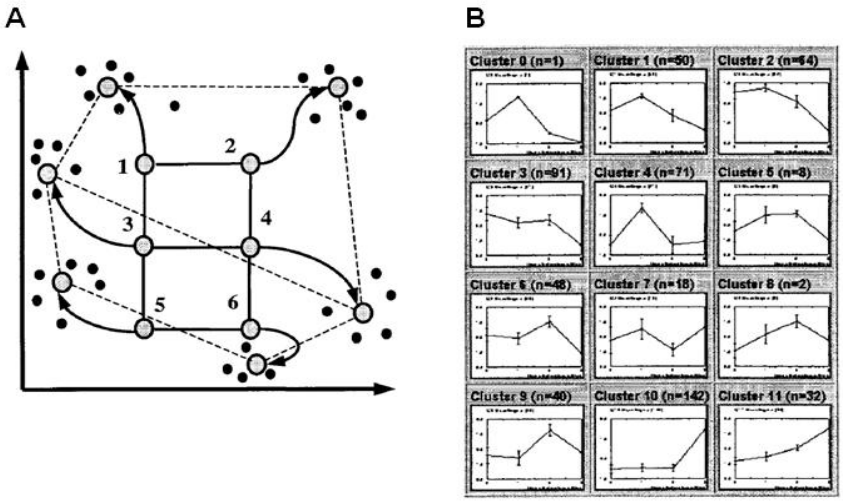

Figure 8: ***The sample structure of the SOM*** *(Tamayo et al., 1999).* ***A)*** *Principle of SOMs. Initial geometry of nodes in 3 3 2 rectangular grid is indicated by solid lines connecting the nodes. Hypothetical trajectories of nodes as they migrate to fit data during successive iterations of SOM algorithm are shown. Data points are represented by black dots, six nodes of SOM by large circles, and trajectories by arrows.* ***B)*** *Example of genechip time course experiment (from 0 to 24 hours) and expression levels of more than 567 genes were grouped by a 4 3 3 SOM.*

Before initiating the analysis, it is necessary to define a geometric configuration for the partitions (2-dimensional rectangular or hexagonal grid), subsequently random vectors are generated for each partition, but before genes can be assigned to partitions, the vectors are first "trained" using an iterative process that continues until convergence so that the data are most effectively separated. At the same time, the models become ordered on the grid so that similar models are close to each other and dissimilar models far from each other. So the order and organization of the nodes (tentative clusters) contain more information than just the actual partition of genes to clusters. In choosing the geometric configuration for the clusters it also specified the number of partitions into which the data is to be divided.

Figure 8B displays the results of the SOM fit. A multi panel figure of plots such as this is one of the most commonly employed display techniques for SOMs. In this experiment, gene expression profiles were measured on cRNA samples prepared from HL-60 cell line cultures at 0, 0.5, 4, and 24 hours after exposure to the phorbol ester PMA. (Tamayo et al., 1999). For each node of the final SOM, an average time course (over the four time points) for the genes mapped to that node was computed. The arrangement of panels in the figure corresponds to the arrangement of nodes on the grid. The line plot displayed in each panel is the average time course for the genes mapped to that node, and the error bars at each time point indicate one standard deviation where the standard deviation is computed from the values recorded at that time point for those genes. For example, the Figure 8B shows that the genes mapping to the nodes in the last row all exhibit expression levels that are relatively constant from baseline (time 0) through 4 hours, but then their expression levels decrease by 24 hours.

SOMs and *k*-means clustering share many of the same advantages as the computational feasibility when large numbers of objects (i.e., genes) are being clustered. In other hand, the principal disadvantage is that it is necessary

to have other source of information, such as PCA, to determine the number of clusters that best represents the available data.

Principal Component Analysis (PCA)

The PCA as previously describe is viewed as a method of reducing data size, then used in the initial step of data filtering. In recent years, is taking up more and more ground using this technique as cluster methods (see example in Figure 9) and then have been implemented software for its application in this area (Peterson, 2003).

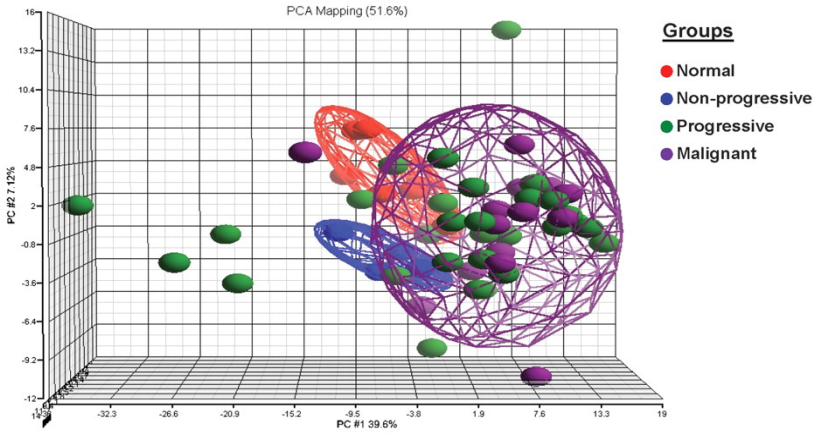

Figure 9: ***Principal Component Analysis****. In this case can be observed that malignant and progressive lesions (purple and green, respectively) and normal and non-progressive leukoplakias (red and blue, respectively) are grouped separately (Cervigne et al., 2009).*

Briefly summarizing, the task of clustering analysis is to find groups of gene with similar expression profile across a number of experiments and/or to find groups of individuals with similar expression profile within population. Clustering methods provide a relatively easy way to organize the cluster information. Together with visualization methods they allow for the user an intuitive way of looking at, understanding and analyzing the data. Different clustering methods and the same method with different premises produce different end results, so the user has to try to find out a useful result. In fact each method and distance has certain properties that can be to emphasize

certain characteristic of the data. Cluster analysis for gene expression data is performed after the original gene expression data is preprocessed. Therefore, the clustering result may be affected by pre-processing procedures such standardization, missing value handling, and flat pattern filtering (D'Haeseleer, 2005). The best way to cluster gene expression data is to use more than one clustering algorithms that may give different results based on different initial conditions should be run several time to find the best solution.

4. GeneChip Technology in Cancer Research

The Affymetrix GeneChip is a popular microarray platform for genome wide expression profiling and has been widely used in functional genomics especially in the classification and characterization of cancers. Indeed, expression microarrays have become a standard tool for cancer study and many microarray data have been generated from different types of cancers (Alon et al., 1999; Chen et al., 2002; Dave et al., 2004; Dhanasekaran et al., 2001; Dyrskjot et al., 2003; Garber et al., 2001; Golub et al., 1999; Irgon et al., 2010; Melis et al., 2010; Meunier et al., 2010; Meyniel et al., 2010; Mikula et al., 2010; Ramaswamy et al., 2003; Saghir et al., 2010; Yamaguchi et al., 2010). Characteristic patterns of gene expression have emerged, reflecting molecular differences between previously known as well as newly discovered characteristics of cancer. Many of the obtained results have been shown to correlate with clinical features, such as survival, prognosis, and treatment sensitivity, as well as traditional histopathological parameters (Gruvberger-Saal et al., 2006).

Gene expression arrays show a great promise by allowing clinicians to stratify cancers by classifying them in subgroups having distinct biological properties and prognoses. For example, the expression of a cohort of several dozen of genes by a tumor may suffice to serve as a strong predictor of its degree of progression or its association with one or another subtype of cancer. Given it, many examples show the application of DNA microarray technology in identifying molecular targets and establishing diagnostic molecular signatures for cancers, here we report two representative works.

A DNA microarray study from Pantel's group showed molecular signature associated with bone marrow micro metastasis in human breast cancer (Woelfle et al., 2003). In this study, gene expression profiles in metastasized breast tumor cells in bone marrow were compared with those in primary tu-

mor cells, and expression analysis showed distinct profiles between these two groups of cells (Figure 10). The differentially expressed genes were related to extracellular matrix remodeling, adhesion, cytoskeleton plasticity, and signal transduction (in particular, the Ras and hypoxia-inducible factor 1 pathways). The array data were confirmed by RT-PCR, which is consistent with immune histochemical analysis of breast tumor tissues. The findings from this study indicate that metastasized breast tumor cells exist as a selective process associated with a specific molecular signature. Study of the functional relevance of this molecular signature will shed light on the molecular diagnosis and therapy of human breast cancer.

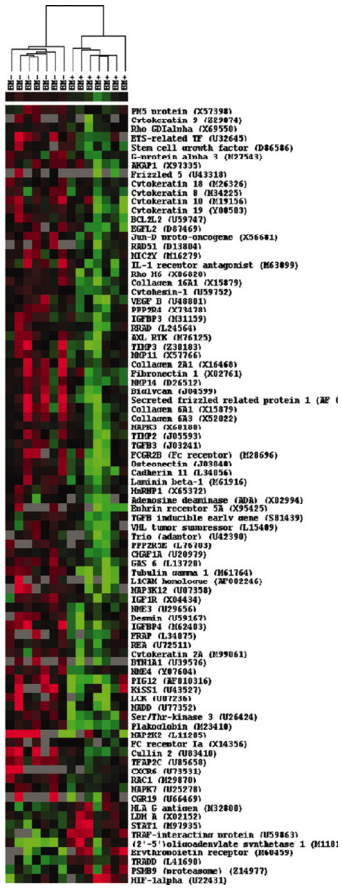

Figure 10: ***Cluster analysis of differentially expressed genes*** *(Woelfle et al., 2003)*

Another example for using DNA microarray technology to study human blood cancers is the molecular classification of human acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) (Golub et al., 1999). In this study, bone marrow mononuclear cells from 11 AML and 27 ALL patients diagnosed pathologically were used as an RNA source for DNA microarray analysis, from which 50 gene predictors that distinguish AML from ALL were derived. These 50 gene predictors were tested and validated on 38 new samples from AML or ALL patients. As a result, 36 of the 38 predictions agreed with the patients' clinical diagnosis (the remaining two were uncertain). This high prediction rate (95%) strongly suggests that DNA microarray technology may be used in the diagnosis of human blood cancers, although the improvement to a 100% prediction rate is the ultimate goal (Figure 11).

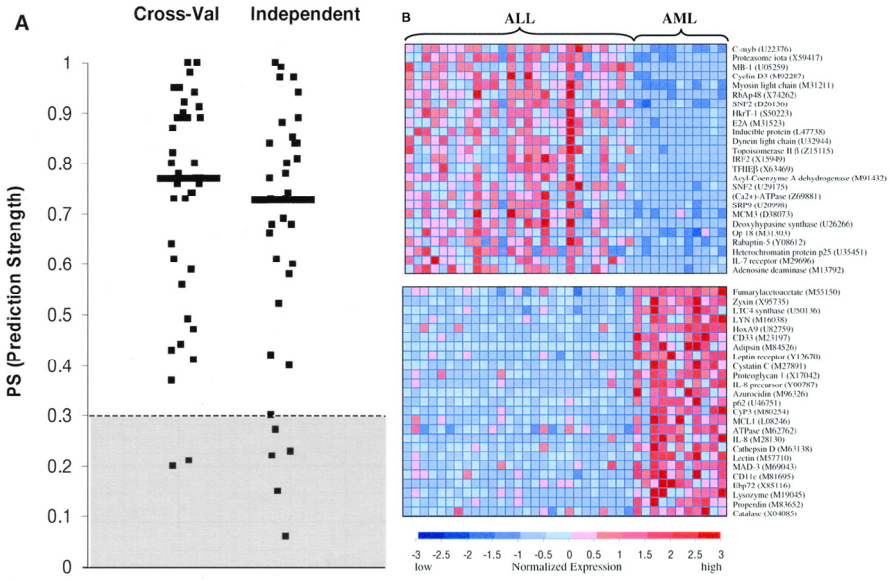

Figure 11: ***Prediction strengths*** *(Gouble et al., 2002).*

Moreover, the recent introduction of specific arrays (including exon-specific arrays and arrays for chromatin immunoprecipitation) together with

technical advancements of established platforms, presents a variety of options to conduct basic as well as translational cancer research. The use of gene expression profiling to address clinical issues clearly illustrates that the molecular signatures of tumors contain information regarding clinical behavior. It is important to remind the fact that transcriptional events only partially correlate with protein levels (Canales et al., 2006; Morey et al., 2006; Rajeevan et al., 2001; Waghray et al., 2001), this means that all the regulatory events downstream gene transcription may need to be taken into account, and also that certain events associated with malignant phenotypes are reflected on the protein levels. However, studies evaluating the overall concordance between protein and RNA expression levels have found wide variability. For example, transcript and protein concordance in the LNCaP prostate cancer cell line has been reported to vary from 32% (Waghray et al., 2001) to 83.5% (Lin et al., 2005). Highly significant correlations in mRNA changes and protein expression levels were found by Orntoft et al. in human carcinomas (Orntoft et al., 2002). Studies such as these suggest that external factors as well as actual biological differences between mRNA and protein abundance might affect the relationships between the two data types.

For accurate study, researchers need to employ a range of different technologies and utilize the whole spectrum of biological and clinical data available in order to elucidate and clarify the complexity of cancer, but the GeneChips are still key analytical tools of functional genomics science.

GeneChip analysis application to cancer knowledge

The Computational methods have become indispensable to biological investigations with the current increase of high-trough up data. Two principal approaches underpin all studies in bioinformatics. First is that of comparing and grouping the data according to biologically meaningful similarities and second, that of analysing one type of data to infer and understand the observations for another type of data. These approaches are reflected in the main aims of the field, which are to understand and organise the information associated with biological molecules on a large scale.

Integration achieves one of the most important imperatives of the new field of the so called Systems Biology, because it reduces the dimensionality of global data needed to deliver useful information about the networks active in the system of interest. The integration of data from different sources provides an effective means to deal with this issue by reinforcing bona fide observations and reducing false negatives. Moreover, because different experimental technologies provide different insights into a system, the integration of multiple data types offers the greatest information about a particular cellular process (Alberghina et al., 2004; Li et al., 2008; Zhu et al., 2007). For example, gene perturbation experiments (i.e., knockouts or RNA interference) and microarrays analysis can reveal relationships between genes that may imply direct physical interactions or indirect logical interactions. Indeed, microarray experiments permit us to look at overall patterns of gene expression in order to understand the architecture of genetic regulatory networks, a global approach that could ultimately lead to complete description of the transcription-control mechanisms in a cell.

Several recent methods have addressed the problem of heterogeneous data integration and network prediction by modeling the noise inherent in high-throughput genomic datasets, especially by using statistical methods, which can significantly improve specificity and sensitivity and allow the robust integration of datasets with heterogeneous properties (Kwoh and Ng,

2007; Srinivasan et al., 2007). Nevertheless, the computational tools necessary to analyse the data are rapidly evolving and no clear consensus exists as to the best method for revealing patterns of gene expression. Indeed, it is becoming increasingly clear that there might never be a "best" approach and that the application of various techniques will allow different aspects of the data to be explored. Furthermore, without a more complete understanding of the underlying biology, particularly of gene regulation, there might never be a single technique that will allow us to find all the relationships in the data. Consequently, choosing the appropriate algorithms for analysis is a crucial element of the experimental design.

Taking into consideration that the development of efficient methods that facilitate the biological interpretation of these data is crucial, in the present work we focus on efficient identification of regulatory mechanisms of cancer, and propose different approaches for analysis and interpretation of gene expression data based on the integration of various types of related biological information and software tools for efficient data analysis.

5. Metabolism and Cancer

Cancer is a disease of uncontrolled cell growth in which cells acquire genetic alterations that allow them to proliferate outside the context of normal tissue development. In the evolution of this transformation, cells acquire mutations that confer selective advantages for the growth of the tumour. Genetic alterations in many of the known oncogenes are selected to adapt cellular metabolism to meet the requirements of rapid cell proliferation as well as autonomous growth and survival in an environment absent of contact with extracellular matrix. In order to divide, a cell needs both to increase its size, and to replicate its DNA, processes that are metabolically demanding, requiring large quantities of proteins, lipids and nucleotides as well as energy. In order to support such large-scale anabolism in rapidly-dividing cancer cells, substantial amounts of metabolic building blocks, particularly glucose and amino acids, must be made available to the tumour. Therefore, tumour cells develop a remarkably different metabolism compared to the normal tissues from which they are derived. The different metabolic requirements of cancer cells from their normal counterparts, involves aberrant activation of most important metabolic pathway as Glycolysis, Lipid biosynthesis, Oxidative Phosphorylation and other important metabolic pathway for cell growth, proliferation and energy production. Accumulating evidence indicates that almost every known oncogene regulates downstream targets that are directly connected to metabolic regulation (Locasale et al., 2009). A detailed biochemical and systems-level understanding of how oncogenes rewire metabolism is essential to understand tumour biology, but concomitantly requires an assessment of the metabolic adaptations required to support the proliferation of cancer cells.

The proliferation and many factors within the tumor microenvironment (as carbon sources) can influence cellular metabolism, resulting in

heterogeneous metabolic activity. Our interest in tumor cells as discussed here involves the metabolic activities that promote their growth and proliferation. In particular, in this chapter we discuss the genetic alterations of metabolic genes and their advantages for the growth of tumor, implemented in two principal works: *Gene expression profiles comparative analysis of immortalized and K-Ras transformed mouse fibroblasts grown in different glucose availability* and *Data recovery and integration from public databases uncovers transformation-specific transcriptional downregulation of cAMP-PKA pathway-encoding genes*.

Increasing attention has been given in recent years to the connection between metabolism and cancer. Under aerobic conditions, normal cells use oxidative phosphorylation as a predominant method for ATP generation. In sharp contrast to normal cells, a common feature to most solid tumors is a major use of glycolysis as ATP source. It has been proposed that the selection of the glycolytic phenotype in cancer cells may be owing to adaptation to hypoxia, a condition characterizing the slowly dividing cancer cells found in large portions of solid tumors not supported by a functional blood supply. Alternatively, the hypoxic adaptation and following glycolytic phenotype could depend on activation of oncogenes (i.e. *ras*), or loss of antioncogenes, mitochondrial dysfunctions.

To better address both causes and consequences of the different metabolism observed in normal and cancer cells, in the first report a detailed transcriptional analysis upon alteration of glucose availability is presented. In particular in this report is specifically addressed the role played by K-Ras oncoprotein in the induction and stabilization of tumor specific metabolic alterations. To this end we have used two cell lines: NIH3T3 mouse immortalized fibroblasts (Normal) and NIH3T3 fibroblasts transformed by an activated form of the K-ras oncogene (Transformed) (Kahn et al., 1987; Yamamoto and Perucho, 1984). (Chiaradonna et al., 2006b). By using these

two cell lines in more recent studies we have shown that the enhanced proliferation potential of transformed cells requires a high initial Glucose (Glc) concentration as well as Glutamine (Gln) in the medium (25 mM Glc and 4 mM Gln, standard cell culture concentrations) since the selective advantage of transformed cells is lost upon growth in sub-optimal Glc (1 mM) or Gln (0.5 mM). Such a requirement correlates with an altered metabolic pattern, as shown by the increased Glc utilization, lactate production, increased expression of glycolytic genes, altered expression of mitochondrial genes, altered mitochondrial morphology, altered activity and reduced capacity to produce ATP of transformed cells (Chiaradonna et al., 2006a; Chiaradonna et al., 2006b). Moreover we could show that in low Glc, transformed cells produce large amount of ROS, due to a specific malfunctioning and assembling of mitochondrial Complex I (Baracca et al., 2010), that may participate to increase the percentage of cell death (Chiaradonna et al., 2006b). Alternatively, reduction of Gln availability in transformed cells causes a strong decrease of their proliferation ability, occurrence largely due to a reduced supply of DNA precursors, as shown by the fact that the growth potential of transformed cells is restored by adding the four deoxyribonucleotides (Gaglio et al., 2009). These responses, consequent to Glc or Gln shortage, are induced by Ras activation, as they are specifically reverted by the expression of the GEF-DN, a Guanine Exchange factor specific for ras that is able to strongly reduce the oncogenic activation of Ras (Bossu et al., 2000) indicating that activation of the Ras pathway strikingly impacts on energy and metabolic aspects of mitochondria functionality (Baracca et al., 2010; Chiaradonna et al., 2006a; Chiaradonna et al., 2006b; Gaglio et al., 2009).

Therefore, to evaluate the relevance of the transcriptional events in such metabolic alterations, time-dependent changes in transcriptome of normal and transformed cells growing in media supplemented with either high (25

mM) or low (1 mM) initial glucose concentration has been performed. Since gene expression profile analysis may help to elucidate the metabolic alterations of cancer cells as well as their inability to react to glucose depletion, we report an overall bioinformatics analysis that indicate a directly link between ras activation and the metabolic alterations of some metabolic pathways in cancer cells.

The second work of this first part of my thesis try to address, at least at transcriptional level, the role of the crosstalk between Ras and cAMP-PKA signaling pathways in the metabolic alterations of cancer cells. In fact the cAMP-PKA signaling pathway is an important regulator of cell fate that controls the activity of metabolic enzymes, transcription factors and cytoskeleton proteins and is strongly associated with the onset of several endocrine and non-endocrine tumors. A fundamental characteristic of cAMP is its ability to stimulate cell proliferation in many cell types while inhibiting in others. Such ability has been related to the fact that cAMP regulates the Ras/Raf/ERK pathway, whose role in cancer onset is well known (about 25% of human cancers have a Ras mutation). Indeed the cAMP pathway is able to suppress ERK signaling through its ability to target C-Raf and conversely, to activate ERK signaling through its ability to target B-Raf (Dumaz and Marais, 2005).

On the other hand cAMP-PKA signaling pathway has an important role in regulation of cellular metabolism. An important role for the metabolic switch observed in cancer cells, has been recognized to mitochondrial dysfunctions impinging their bioenergetics activity, due to mitochondrial DNA mutations, altered mitochondrial enzymes expression and more recently also alterations of respiratory chain complexes composition, stability and activity caused by post-translational events, i.e., phosphorylation, of mitochondrial proteins (Alchanati et al., 2006; Isidoro et al., 2004; Lee and Wei, 2009; Lu et al., 2009), given that these modifications

appear to be essential for physiological regulation of mitochondria activity (reviewed in Regulation of mitochondrial oxidative phosphorylation by second messenger-mediated signal transduction mechanisms. Boneh A. Cell. Mol. Life Sci. 63, 2006). Moreover, numerous findings have been collected indicating that such mitochondrial alterations do play a relevant role in the etiology of cancer and in the appearance of an altered glycolysis in cancer cells and tumor tissues. However, only few reports identified a strict association between metabolic changes in cancer cells and mitochondrial complexes composition and activity (Simonnet et al., 2003). Recently, it has been shown that proteins belonging to PKA pathway are localized both on the external membrane and inside of mitochondria, where they are able to control mitochondrial activity by a phosphorylation mechanism (Acin-Perez et al., 2009). Moreover, it has been shown that activation *in vivo* of the cAMP/PKA cascade promotes Complex I activity, decreases ROS cellular levels, regulates mitochondrial morphology and activity and modulates, in cell specific manner, the apoptotic process by specific phosphorylation of apoptotic proteins (Bellis et al., 2009; Chang and Blackstone, 2007; Cribbs and Strack, 2007; Harada et al., 1999; Piccoli et al., 2006; Robinson-White et al., 2009). Connection between PKA pathway and cell transformation has been widely observed (Tortora and Ciardiello, 2002) and recently such a connection has been also demonstrated in terms of transcriptional regulation.

Therefore, the comparison between the transcriptional profiles of PKA related pathway encoding genes, recovered from several normal tissues and from the NCI60 transformed cells collection (Ross et al., 2000), can be useful for understanding the activity of oncogenic Ras on the transmission signal activated by cAMP-PKA axis (Balestrieri et al., 2009; Chiaradonna et al., 2008).

5.1 Gene expression profiles comparative analysis of immortalized and K-Ras transformed mouse fibroblasts grown in different glucose availability.

Increasing attention has been given in recent years to the connection between metabolic alterations and cancer (DeBerardinis et al., 2008). Under aerobic conditions, normal cells use oxidative phosphorylation as a predominant source for ATP generation. In sharp contrast to normal cells, a common feature of most cancer cells is a major use of glycolysis to produce ATP (Mazurek and Eigenbrodt, 2003; McFate et al., 2008; Ramanathan et al., 2005; Warburg, 1956). It has been proposed that the selection of the glycolytic phenotype in cancer cells may be owing to adaptation to hypoxia. Alternatively, the hypoxic adaptation and following glycolytic phenotype could depend on activation of oncogenes (i.e., ras), or loss of antioncogenes given the fact that the consequences of glucose deprivation have been extensively described in several cancer cells and tissues, for more detail see (Dang and Semenza, 1999; Gatenby and Gillies, 2004).

In order to critically analyze the molecular basis of the change of carbon metabolism in cancer cells, we compared the transcriptional profiles of normal and transformed cells grown in 25 mM glucose (normal cell culture condition) and 1 mM glucose (hypoglycemic condition) along a time course of 72 hours. NIH3T3 cells are a genetically well defined immortalized cell line that has long established as a model parental cell line for the study of cell transformation (Yamamoto and Perucho, 1984). Ras proteins are intracellular switches whose activation state (i.e., their binding to GDP or GTP) controls downstream pathways leading to cell growth and differentiation. The activation state of Ras proteins is governed through the competing action of GTPase Activating Proteins (GAP) and Guanine nucleotide Exchange Factors (GEF). Mutation of the ras gene, identified in about 25% of all human tumors, is a critical event in the onset of different

malignant phenotypes. Also deregulation of either GAP or GEF activity may result in hypo- or hyper-activation of downstream pathway(s), so that for instance over-expression of a GEF or inactivation of a GAP may both result in cell transformation (Dupuy et al., 2001; Vogel et al., 1999).

5.1.1 Results

Selection of differentially expressed genes

We used mouse normal NIH3T3 mouse fibroblasts (normal cells) and NIH3T3 cells transformed by an activated form of the K-ras oncogene (transformed cells). Total RNA was harvested from cells cultured in two different concentration of glucose for various times, in particular cells grown in 25mM glucose (standard culture condition) and 1mM glucose (suboptimal culture condition) during a time course 0, 24, 48 and 72 hours (see experimental design in Figure 12). Labeled probes synthesized from cellular mRNA were hybridized to oligonucleotide microarray (array Mouse 403 2.0) that detect the expression of more than 45000 probe sets representing over 34000 well-substantiated mouse genes. Gene expression data were subjected to normalization and filtering procedures as described in Methods. We obtained a list of ~ 20000 probe sets, all probe sets were used as input for the algorithm of screening and a list of statistically well-characterized 9351 unique genes was obtained, and called *working list* (see details in Methods).

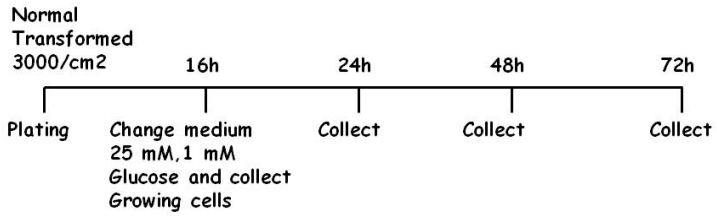

Figure 12: ***Experimental design.***

The differentially expressed genes of the working list were identified by Welch one-way ANOVA analysis. P-values were calculated for each gene over the time courses at time 0, 24, 48 and 72 hours. The analyses were performed on the log_2-fod change in gene expression for time *n* versus time 0, for every time course separately. To reduce the detection of false positives, the p-values were adjusted by Benjamin and Hochberg false discovery rate method (MTC) by using a cut-off of 0.001 and a cut-off level of 1.5 fold increase or decrease of time *n* versus time 0. This procedure produced a list of 1210 genes shown in hierarchical cluster in Figure 13.

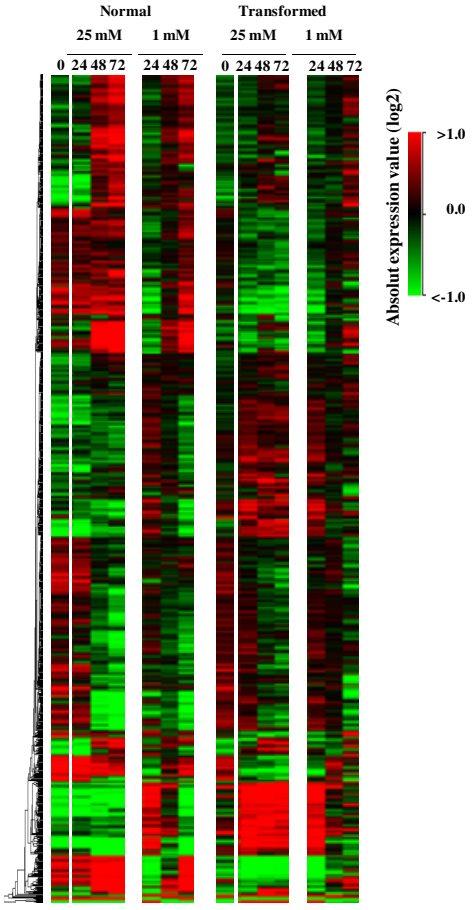

Figure 13: ***Unsupervised hierarchical clustering of 1210 unique genes selected with Welch's ANOVA.*** *The heat red color indicates higher expression and green color indicates lower expression. Hierarchical*

clustering of the expression levels was performed by using the Euclidean measure for the similarity and the centroid linkage for clustering.

Enrichment analysis

Ontological and KEGG pathway analysis of the differentially regulated genes over the time were performed using GeneCodis tool (http://genecodis.dacya.ucm.es/). The program computes a significance significant p-value using hypergeometric test and FDR method with a cut-off of 0.05 (see Methods). Gene ontology analysis focusing on the biological processes recognized by the 1210 genes, identified several processes including Transcription Regulation, Transport, Multicellular Organismal Development, Cell Cycle as well as Metabolic Processes (Table 1). In particular in Table 1 we reported a list of 13 statistically biological processes identified and ranked in order to the decrease of number genes present in the class (N° genes). Moreover in the Table 1 also have been reported the p-value and the correct p-value (FDR method). In particular the Metabolic processes comprised a total of 483 genes of which 46 genes were modulated.

Table 1: ***Functional classification of significantly expressed genes over the four time courses.***

N° genes	p-value	Correct p-value	Biological Process (GO term)
120	1.77E-11	2.48E-10	transcription
109	5.80E-10	5.42E-09	transport
64	5.32E-08	3.72E-07	multicellular organismal development
50	6.40E-12	1.79E-10	cell cycle
46	9.18E-06	3.67E-05	metabolic process
41	9.43E-07	5.28E-06	protein transport
35	1.13E-03	3.52E-03	ion transport
31	6.51E-03	1.82E-02	cell differentiation
15	1.02E-03	3.56E-03	carbohydrate metabolic process
14	1.57E-06	7.32E-06	response to stress
5	1.91E-02	4.11E-02	cytoskeleton organization
4	9.53E-03	2.43E-02	cell death
3	1.66E-02	3.88E-02	cellular component organization

To further detail the analysis, the identification of regulated pathways was performed. In particular, by using GeneCodis tool that collect information from **K**yoto **E**ncyclopedia of **G**enes and **G**enomes (KEGG) database, we identified the most important KEGG pathways (41 pathways) identified in our working list (Figure 14). In the Figure 14 the pathways are represented in accord increased p-value (the first x-axis, maximum p-value of 0.05), and the number of genes (the second x-axis, maximum value 35) for each pathway (y-axis). The results showed that among these pathways, several metabolic pathways were altered in transformed cells, as i.e., Glutathione metabolism, Sulfur and Fructose and Mannose metabolism, further confirming the relevant role of metabolic alterations in K-ras dependent transformation. However several other statistical relevant pathways were identified, including one called “Pathway in cancer”.

The Pathway in cancer comprises 321 protein-encoding mouse genes that are already shown to be implicated in oncogenesis (http://www.genome.jp/kegg/pathway.html). Among these genes, 32 are modulated in their expression along the time course analyzed (10% of genes belonging to the pathway). This pathway includes different cancer-related pathways as Cell cycle, Wnt signaling pathway, ECM receptor, MAPK signaling pathway, Apoptosis, ect. We have to underline that most of these pathways were already identified in the previous KEGG analysis with a significant of p-value (Figure 14).

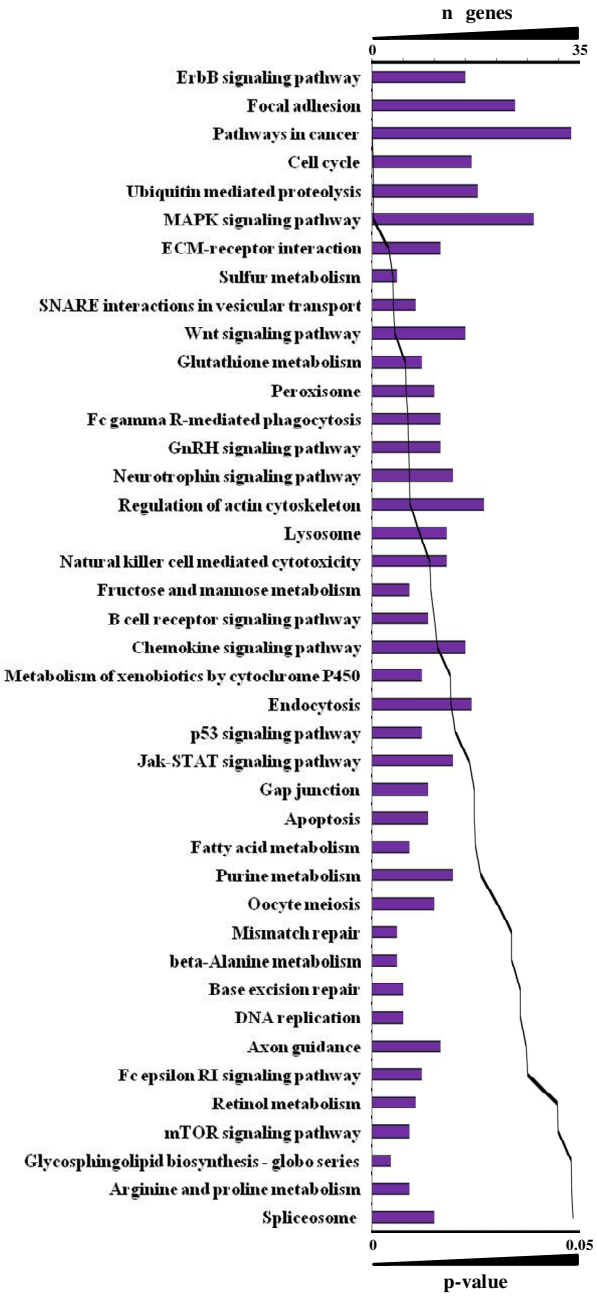

Figure 14: ***KEGG significantly altered over the time.*** *KEGG pathways were identified as significantly altered by using a hypergeometric test and FDR with a p-value cut-off of 0.05. The first x-axis represent the p-value and the second x-axis represent the number of genes for each KEGG pathway shown in y-axis.*

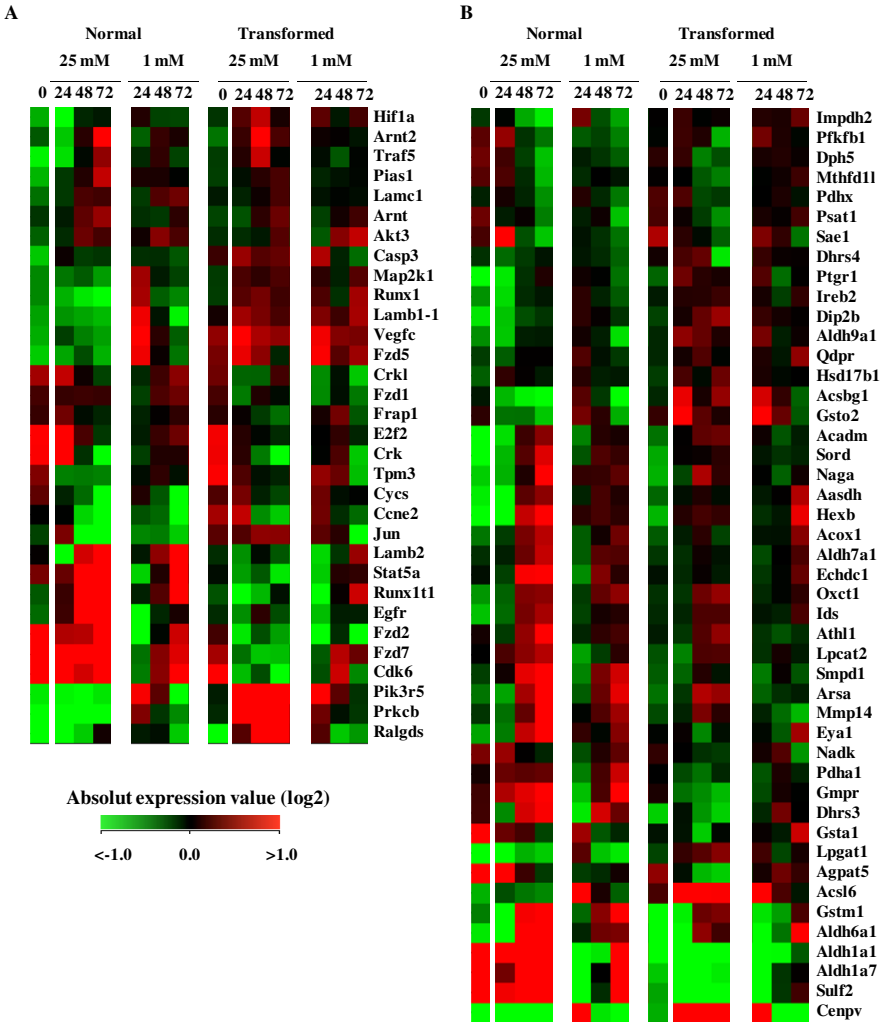

Figure 15: ***Hierarchical clusters of the genes present in A) Pathway in cancer and B) Metabolism processes.*** *Genes are represented on rows and the experimental samples are represented on columns. Genes with high expression levels are in red, intermediate-level expression in black and low-level expression in green. In the bottom of the figure is shown the color scale used to represent the log_2 transformed values.*

In Figure 15 has been shown the Heat map of 32 regulated genes in the *Pathway in Cancer* (Figure 15A) and 46 regulated genes in *Metabolic Process* (Figure 15B), the complete list of genes and their description are reported in Tables 2 and 3 respectively.

Table 2: ***Genes involved in Pathway in Cancer (32 genes.)***

Gene Symbol	**Description**
Hif1a	hypoxia inducible factor 1, alpha subunit
Arnt2	aryl hydrocarbon receptor nuclear translocator 2
Traf5	TNF receptor-associated factor 5
Pias1	protein inhibitor of activated STAT 1
Lamc1	laminin, gamma 1
Arnt	aryl hydrocarbon receptor nuclear translocator
Akt3	thymoma viral proto-oncogene 3
Casp3	caspase 3
Map2k1	mitogen-activated protein kinase kinase 1
Runx1	runt related transcription factor 1
Lamb1-1	laminin B1 subunit 1
Vegfc	vascular endothelial growth factor C
Fzd5	frizzled homolog 5 (Drosophila)
Crkl	v-crk sarcoma virus CT10 oncogene homolog (avian)-like
Fzd1	frizzled homolog 1 (Drosophila)
Frap1	FK506 binding protein 12-rapamycin associated protein 1
E2f2	E2F transcription factor 2
Crk	v-crk sarcoma virus CT10 oncogene homolog (avian)
Tpm3	tropomyosin 3, gamma
Cycs	cytochrome c, somatic
Ccne2	cyclin E2
Jun	Jun oncogene
Lamb2	laminin, beta 2
Stat5a	signal transducer and activator of transcription 5A
Runx1t1	runt-related transcription factor 1
Egfr	epidermal growth factor receptor
Fzd2	frizzled homolog 2 (Drosophila)
Fzd7	frizzled homolog 7 (Drosophila)
Cdk6	cyclin-dependent kinase 6
Pik3r5	phosphoinositide-3-kinase, regulatory subunit 5, p101
Prkcb	protein kinase C, beta
Ralgds	ral guanine nucleotide dissociation stimulator

Table 3: ***Genes involved in Metabolism process (46 genes).***

Gene Symbol	Description
Impdh2	inosine 5'-phosphate dehydrogenase 2
Pfkfb1	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1
Dph5	DPH5 homolog (S. cerevisiae)
Mthfd1l	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1-like
Pdhx	pyruvate dehydrogenase complex, component X
Psat1	phosphoserine aminotransferase 1
Sae1	SUMO1 activating enzyme subunit 1
Dhrs4	dehydrogenase/reductase (SDR family) member 4
Ptgr1	prostaglandin reductase 1
Ireb2	iron responsive element binding protein 2
Dip2b	DIP2 disco-interacting protein 2 homolog B (Drosophila)
Aldh9a1	aldehyde dehydrogenase 9, subfamily A1
Qdpr	quinoid dihydropteridine reductase
Hsd17b1	hydroxysteroid (17-beta) dehydrogenase 1
Acsbg1	acyl-CoA synthetase bubblegum family member 1
Gsto2	glutathione S-transferase omega 2
Acadm	acyl-Coenzyme A dehydrogenase, medium chain
Sord	sorbitol dehydrogenase
Naga	N-acetyl galactosaminidase, alpha
Aasdh	aminoadipate-semialdehyde dehydrogenase
Hexb	hexosaminidase B
Acox1	acyl-Coenzyme A oxidase 1, palmitoyl
Aldh7a1	aldehyde dehydrogenase family 7, member A1
Echdc1	enoyl Coenzyme A hydratase domain containing 1
Oxct1	3-oxoacid CoA transferase 1
Ids	iduronate 2-sulfatase
Athl1	ATH1, acid trehalase-like 1 (yeast)
Lpcat2	lysophosphatidylcholine acyltransferase 2
Smpd1	sphingomyelin phosphodiesterase 1, acid lysosomal
Arsa	arylsulfatase A
Mmp14	matrix metallopeptidase 14 (membrane-inserted)
Eya1	eyes absent 1 homolog (Drosophila)
Nadk	NAD kinase
Pdha1	pyruvate dehydrogenase E1 alpha 1
Gmpr	guanosine monophosphate reductase

Dhrs3	dehydrogenase/reductase (SDR family) member 3
Gsta1	glutathione S-transferase, alpha 1 (Ya)
Lpgat1	lysophosphatidylglycerol acyltransferase 1
Agpat5	1-acylglycerol-3-phosphate O-acyltransferase 5
Acsl6	acyl-CoA synthetase long-chain family member 6
Gstm1	glutathione S-transferase, mu 1
Aldh6a1	aldehyde dehydrogenase family 6, subfamily A1
Aldh1a1	aldehyde dehydrogenase family 1, subfamily A1
Aldh1a7	aldehyde dehydrogenase family 1, subfamily A7
Sulf2	sulfatase 2
Cenpv	centromere protein V

Protein-Protein Interaction analysis

Another level of investigation involves a Protein-Protein Interactions map (PPI's) by using the gene expression data. In this regard, to extract gene related information from literature, the GeneSpring Natural language processing algorithm was used. In particular different type of relations (interaction between entities) have been analyzed from GeneSpring database, like binding, expression, metabolism and each relation has been considered significant if characterized by two or more entities or nodes (proteins). In this kind of analysis, the *degree of a node*, that give the measurement of its role in the network, can be defined as the number of its non-redundant interactions with other proteins in the network. Proteins with high degree are central proteins (or hub) and are critical to the integrity of the network. As shown in Figure 16A, the network obtained applying the GeneSpring algorithm was a layout of 248 proteins (node) and 519 PPI's (interaction). The top central proteins with high degree are AHR, RUNX2, CCL2, MAP3K1, EGFR, JUN, KRT14 and EDN1. The Figure 16B also shows the change in expression profile of hub genes.

Notable several of these genes show opposite behavior in the different conditions compared. Examples of these behaviors are the AHR gene (low values of expression in normal cells and high value of expression in

transformed cells, both grown in 25 mM glucose), and RUNX2 and KRT14 genes (high expression in normal cells and low expression in transformed cells, both grown in 25 mM glucose).

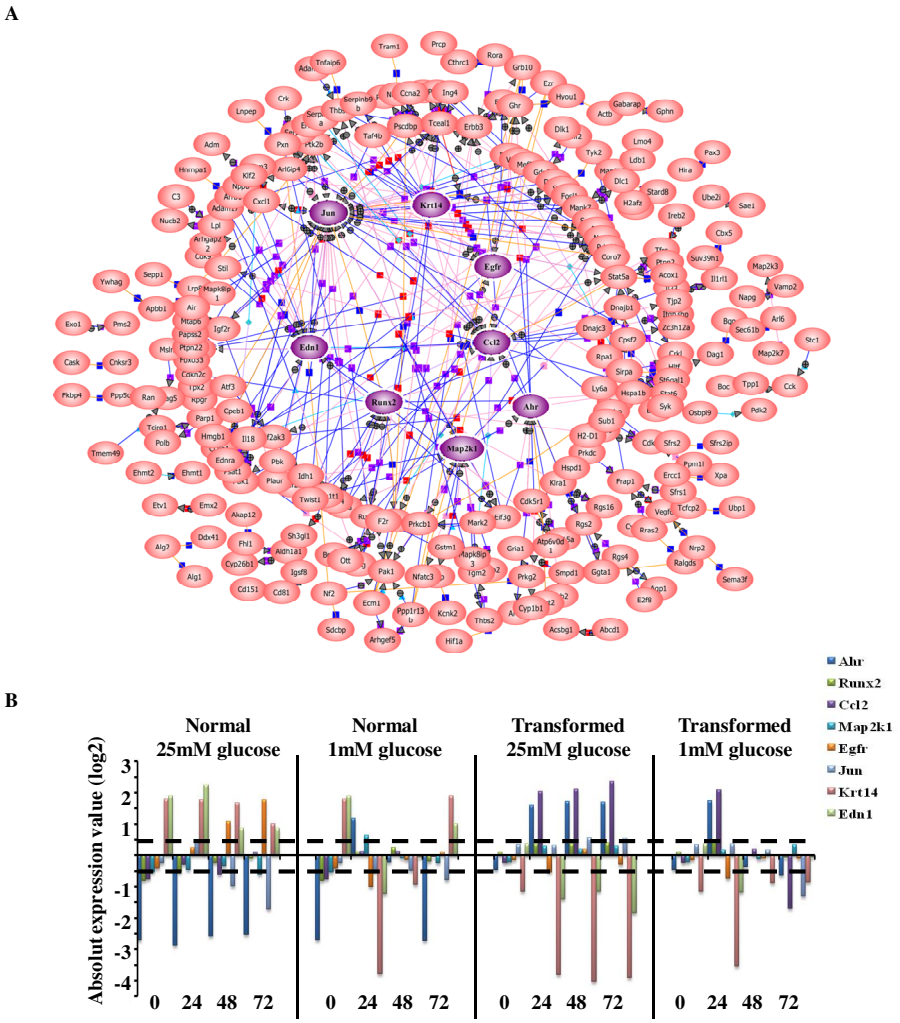

Figure 16: ***Protein-Protein interaction network. A)*** *the layout network of 248 proteins and 519 PPI's. The principal hubs are colored in purple;* ***B)*** *Expression profiles of 8 principal hub of networks.*

Moreover other genes shown a individual time-dependent regulation depending on condition analyzed. In example EGFR gene changes its level of expression only in normal cells at 48 and 72h (expression between 1 and

1.5). Another example is the oncogene JUN that has a low level of expression at time 72h of normal and transformed cells grown in 25mM glucose and in transformed cells grown in 1mM.

5.1.2 Discussion

In this study we address, in normal and transformed mouse fibroblasts, the relation between nutrient-sensitive signaling pathways and cancer. By comparing the transcriptional profiles regulated by the oncogene and glucose, we have shown that a significant number of pathways are deregulated at transcriptional level, in particular we have observed pathways involved in Cellular Metabolism, Cell Cycle, Signal Transduction and Apoptosis, that are pathways well-known related to cancer and hence components of more global macro-pathway called Pathway in Cancer. In the future the role of some of these genes in both positive and negative regulation of the pathways to which they belong will be investigated. Interestingly, as briefly described above, the 8 hubs identified in this study are known to have a close relation with the tumorigenesis.

The oncogene *JUN* is the putative transforming gene of avian sarcoma virus 17. It encodes a protein highly similar to the viral protein that interacts directly with specific target DNA sequences to regulate gene expression. Jun is believed to be an oncogene (Vogt and Bos, 1990).

The aryl hydrocarbon receptor (*AHR*), a ligand-activated member of the basic-region helix-loop-helix/period-aryl hydrocarbon nuclear translocator-simple-minded (bHLH/PAS) family of transcription factors, controls a variety of developmental and physiologic events, including induction of drug metabolizing enzymes, xenobiotic detoxification, neurogenesis, tracheal and salivary duct formation, circadian rhythms, response to hypoxia, and hormone receptor function (Pocar et al., 2005). Historically, studies of AhR pathway have focused on the transcriptional regulation of genes encoding xenobiotic metabolizing enzymes such as cytochrome P450 enzymes

(Hillegass et al., 2006). Recent studies demonstrated a close relation between AhR and mammary gland tumorigenesis (Marlowe and Puga, 2005; Schlezinger et al., 2006). The AhR gene polymorphisms have been linked to an increased risk of lung and breast cancers (Kim et al., 2007; Long et al., 2006). Increased expression of AhR has been reported in lung, breast, and pancreatic cancers in humans (Lin et al., 2003; Schlezinger et al., 2006). These data indicated a close relationship between AhR and tumorigenesis. However, the relation between AhR and tumor progression is not yet completely clear.

RUNX2 gene is a member of the RUNX family of transcription factors and encodes a nuclear protein with a Runt DNA-binding domain. In particular Runx2 is a master regulator of bone formation, reviewed in (Barnes et al., 2003; Pratap et al., 2006), and it is abnormally and highly expressed in MDA-MB-231 breast cancer cells that metastasize to bone and form osteolytic lesions (Pratap et al., 2008). Notably, Runx2 is not significantly detected in normal breast or prostate epithelial cells (Barnes et al., 2004a; Inman and Shore, 2003). Runx2 functions in many regulatory processes of osteoblast including epigenetic control of genes during mitosis (Young et al., 2007), suppression of cell growth (Pratap et al., 2003), cellular senescence (Zaidi et al., 2007) and bone turnover. A unique property of Runx2 is its sub nuclear targeting to foci recruiting co-regulatory factors that mediate transduction of Wnt, Src, BMP and TGFβ signaling. These pathways are activated in tumor cells (Hall et al., 2006; Kingsley et al., 2007).

Monocyte chemoattractant protein 1 (*CCL2*) is a member of the CC chemokine family and is known to promote monocyte chemotaxis to sites of inflammation. CCL2 is an important determinant of macrophage and monocyte infiltration in breast, cervix, and pancreatic carcinomas (Balkwill and Mantovani, 2001). Other studies have demonstrated that CCL2 localizes

to tumor epithelial cells (Negus et al., 1995), and that the levels of CCL2 expression have been correlated with the involvement of lymphocyte and macrophage localization in secondary sites of tumor formation (Negus et al., 1997). Other findings suggest that CCL2 may act directly on the epithelial cells of several human carcinomas and may regulate the migration and invasive properties of tumor cells, resulting in enhanced metastatic potential.

Mitogen activated protein kinase kinase kinase (*MAP3K1*) is a serine/threonine protein kinase of the MAP3K super family. Compelling evidence indicates that MAP3K1 is involved in the regulation of diverse functions in a tissue- and cell-type-specific manner. Most effects of MAP3K1 depend on its kinase activity, which upon induction by upstream cues leads to the phosphorylation and activation of the MAP2K-MAPK cascades. In most cases, MAP3K1 preferentially activates MAP2K4 and MAP2K7, which are upstream activators for the c-Jun N-terminal Kinases (JNKs) and/or p38 MAPKs (Davis, 2000; Xia et al., 2000; Xia et al., 1998), but also can regulate the extracellular signal-regulated kinase (ERK) MAPKs Mek1 and Mek2 (Karandikar et al., 2000; Lu et al., 2002; Witowsky and Johnson, 2003). Furthermore, some reports suggest that MAP3K1 can phosphorylate the IκB kinases IKKα and IKKβ and induce activation of nuclear factor-κB (NF-κB) downstream of tumor necrosis factor-α (TNF-α) (Lee et al., 1997). Moreover, MAP3K1 might participate in several pathological processes. In fact, is required for Bcr-Abl-induced STAT3 activation and LIF-independent self-renewal, suggesting a role in Bcr-Abl-mediated leukemogenesis (Nakamura et al., 2005); it is required for the induction of cardiac hypertrophy (Minamino et al., 2002); and finally, in mammary glands, MAP3K1 contributes to polyoma middle T antigen (PyMT)-mediated primary mammary tumor cell dissemination and lung metastasis (Cuevas et al., 2006). Moreover, MAP3K1 plays also role in anti-apoptotic and pro-apoptotic process (Faris et al., 1998; Yujiri et al., 1998).

Apart from its role in the MAPK signaling, MAP3K1 has been found to directly interact with and/or phosphorylate transcription factors, such as STAT3, and co-factors, such as transducer of regulated CREB activity 1 (TORC1) and p300 (Siu et al., 2008).

Endothelin-1 (*EDN1*) is a growth factor and plays a key role in cell growth and differentiation, as well as in vascular homeostasis in mammals. EDN1 is primarily expressed in vascular epithelial cells where it plays an important role in maintaining proper vascular tone (Kedzierski and Yanagisawa, 2001). EDN1 is frequently secreted by many solid tumors, including prostate, colorectal, liver, breast and ovarian cancers (Kusuhara et al., 1990). In these tumor cells, EDN1 has been shown to promote cell proliferation, suppress apoptosis, promote metastasis, and facilitate angiogenesis (Bagnato and Spinella, 2003). EDN1 exerts its diverse functions through two cell-surface receptors, EDNRA and EDNRB (Kedzierski and Yanagisawa, 2001) In animal tumor models, endothelin receptor antagonists have demonstrated remarkable effects in suppressing tumor growth (Bagnato et al., 2002). In clinical trials, these antagonists significantly delay the progression of metastasis in hormonal-refractory prostate cancers. These observations implicate abnormal expression of EDN1 to be a key step in tumorigenesis of many solid tumors.

The epidermal growth factor receptor (*EGFR*) regulates the intracellular effects of ligands such as EGF and transforming growth factor-α (TGFα) (Carpenter and Cohen, 1990; Wells, 1999; Yarden and Sliwkowski, 2001). EGFR regulates a number of cellular functions, including proliferation and survival, that are also crucial in tumorigenesis, thus making EGFR a promising target for the cancer therapies (Jorissen et al., 2003).

KRT14 gene encodes a member of the keratin family, the most diverse group of intermediate filaments. This gene product, a type I keratin, is usually found as a heterotetramer with two keratin 5 molecules, a type II

keratin. Together they form the cytoskeleton of epithelial cells. Dominant mutations in the genes encoding these proteins were shown to disrupt the keratin filament cytoskeleton (Corden and McLean, 1996).

Integration of results presented in this work is an example of genome-wide computational approaches that can be able to lead to a system-level understanding of the links between K-Ras-induced transformation and carbon metabolism.

5.2 Data recovery and integration from public databases uncovers transformation-specific transcriptional downregulation of cAMP-PKA pathway-encoding genes.

As previously described, the Ras pathway is able to crosstalk with the cAMP-PKA pathway by some typical signal transduction mechanisms (i.e. protein-protein interaction, protein phosphorylation). Moreover, through its ability to regulate the activity of a large number of transcription factors (Bader et al., 2005; Treisman, 1996), the Ras pathway is able to control several transcriptional programs leading to proliferation, differentiation, metabolism, cytoskeletal reorganization and immune response. Such transcriptional programs are the result of ras-specific effectors stimulation (Malumbres and Barbacid, 2003). Until now more than ten distinct functional classes of proteins have been involved as effectors of the small GTPase Ras, but the best studied are Raf kinases, type I phosphoinositide (PI) 3-kinases, Ral-guanine nucleotide exchange factors (Ral-GEFs), the Rac exchange factor Tiam1, and phospholipase C (Downward, 2003).

Raf and phosphatidylinositol 3-kinase (PI3K) were the first two identified Ras effectors and the main focus of research investigating Ras functions (Marais et al., 1998). Raf promotes cell proliferation and differentiation through the MAP kinase (MAPK) pathway (McKay and Morrison, 2007), at the same time as PI3K generates anti-apoptotic signaling, directly or through Akt pathway activation (Anderson et al.,

1999). Both signaling pathways can activate two different signals distinct for their response timing. Indeed both MAPK and PI3K are able to activate phosphorylation cascades that lead, as primary effect, to post-translational modification of several substrates (membrane targets, cytosolic targets, cytoskeletal targets and nuclear targets), which rapidly activate functional processes. Early response to Ras signaling is quite fast: for instance in resting cells stimulated with mitogens, Ras-GTP level increases within 2 minutes from stimulation with serum (Marais et al., 1998). Raf-1 undergoes transient activation within 2-3 minutes, and rapidly activates the mitogen-activated protein kinase (MAPK) cascade whose most downstream component, ERK, rapidly moves into the nucleus. Here it phosphorylates nuclear proteins, notably transcription factors (Davis, 1995) whose activity can be controlled by regulating their sub-cellular localization, expression, stability, ability to bind to other components of transcriptional complexes and to DNA, and their ability to remodel chromatin structure (Hazzalin and Mahadevan, 2002). Transcription factors are under the control of MAPK pathway, including members of the ETS family (i.e. Ets-1, Ets-2, PU-1), MADS box family (i.e. MEF2A, MEF2C, Sp1), Zinc Finger family (i.e. GATA-2 and GATA-4), bZip family (i.e. Fra-1, c-Jun, JunB, JunD, ATF-2, c-Fos and CREB), bHLH family (i.e. c-Myc, MITF), Nuclear Hormone Receptor (i.e. PR, GR and ER) as well as other transcription factors (i.e., SMAD1, STAT1) and co-regulatory proteins (i.e., CBP, p300) (Davis, 1995; Treisman, 1996).

Like ERKs, Akt and other targets of PI3K signaling can phosphorylate and activate transcription factors (Chang et al., 2003). Akt protein can control several transcription factors directly or indirectly. Direct targets are the forkhead box proteins, FOXO, and the cell cycle inhibitor, MIZ1, which are both inhibited upon AKT-mediated phosphorylation. AKT-dependent

regulation of p53, nuclear factor B (NFkB), c-MYC, activator protein 1 (AP1) and beta-catenin is indirect (Bader et al., 2005).

Such an observation led us to re-analyze, by using a generalized workflow for data recovery and integration, available data from multiple global assays and several databases (genomics, transcriptomics, promoter analysis and literature). In particular we searched for information for genes encoding proteins of the downstream branch of the PKA pathway (starting from adenylyl cyclase and downstream) in tumor cell lines (NCI60 cells) as a function of mutational activation of different pathways (notably the Ras and PI3K pathway) in comparison with the corresponding normal tissues, with the aim to define better the connection between these pathways in cancer cells (Oda et al., 2005).

The integration of data from multiple genome-wide assays is essential for understanding dynamic spatio-temporal interactions within cells. Such integration, which leads to a more complete view of cellular processes, offers the opportunity to rationalize better the high amount of "omics" data freely available in several public databases. In particular, integration of microarray-derived transcriptome data with other high-throughput analyses (genomic and mutational analysis, promoter analysis) may allow us to unravel transcriptional regulatory networks under a variety of physiopathological situations, such as the alteration in the cross-talk between signal transduction pathways in transformed cells (Balestrieri et al., 2009).

Here we sequentially apply web-based and statistical tools to a case study: the role of oncogenic activation of different signal transduction pathways in the transcriptional regulation of genes encoding proteins involved in the cAMP-PKA pathway. Through its ability to regulate the activity of a large number of transcription factors, the Ras pathway is able to control several transcriptional programs leading to proliferation, differentiation, metabolism, cytoskeletal reorganization and immune response. Cyclic AMP

(cAMP) is a ubiquitous intracellular second messenger whose major intracellular target in eukaryotes is protein kinase A (PKA). Wide evidence for cross talk between the Ras and cAMP-PKA pathways is available. After reviewing some features of Ras and PKA signaling that are relevant for cancer biology, we re-analyze available genome-wide expression data for genes encoding proteins of the downstream branch of the PKA pathway in human tumor cell lines as a function of the mutational state of the Ras pathway. The observed Ras-dependent pattern of regulation of the analyzed genes may contribute to explain how the cAMP/PKA axis is involved in oncogenic processes induced by Ras (Chiaradonna et al., 2008).

Genome-wide, large-scale "omics" experimental technologies give different, complementary perspectives on the structure and regulatory properties of complex systems. Even the relatively simple, integrated workflow presented here offers opportunities not only for filtering data noise intrinsic in high throughput data, but also to progressively extract novel information that would have remained hidden otherwise. In fact we have been able to detect a strong transcriptional repression of genes encoding proteins of cAMP/PKA pathway in cancer cells of different genetic origins. The basic workflow presented herein may be easily extended by incorporating other tools and can be applied even by researchers with poor bioinformatics skills (Balestrieri et al., 2009).

5.2.1 Results

Gene-expression profiling has been applied extensively in cancer research. As a first step to identify regulatory mechanisms underlining gene-expression profiles it is necessary to extract, filter, cross-reference and structure information from cancer-related data sets (Lander, 1999). The aim of this work has been the identification of cancer-specific specific gene expression signatures in genes encoding proteins involved in the cAMP-PKA pathway. In particular we wished to identify, if present, differences

between primary normal tissues and cancer cells and search for correlation with the pathway mutationally activated in any given transformed cell line by integrating an accurate analysis of recovered data from several databases with the application of different statistical tests.

Transformation-dependent, transcriptional remodeling of the PKA pathway encoding genes in 60 human cancer cell lines (NCI60)

The NCI60 cell collection includes cell lines derived from colorectal, renal, ovarian, breast, prostate, lung and central nervous system cancers, as well as leukaemias and melanomas (Table 4), that are most commonly used in cancer research and drug screening (Wang et al., 2006). A good correlation between transcriptional profiles of the cell lines and their tumor cancer of origin (Ross et al., 2000; Wang et al., 2006) has been found for 51 out of 59 cell lines. NCI60 transcriptional profiles are available in public databases.

Table 4: ***The cancer cell lines in the NCI60 collection sorted by tissue of origin.***

Tumor Type	**Cell lines**
Breast	HS578T, MDA-MB231, MD-MB435, MCF7, T47, MDA-N, BT549
CNS	SF295, SF359, SNB19, U251, SF268, SNB75
Colon	HCC2998, HCT116, HCT15, SW620, COLO205, HT29, KM12
Leukaemia	CCRF-CEM, RPMI-8226, HL60, MOLT4, K562, SR
Melanoma	SK-MEL2, LOXIMVI, M14, MALME-3M, SK-MEL28, SK-MEL5, UACC257, UACC62
Lung	A549, HOP62, NCI-H23, NCI-H460, EKVX, NCI-H226, NCI-H322, NCI-H522, HOP92
Ovarian	OVCAR5, OVCAR8, SKOW3, IGROV1, OVCAR3, OVCAR4
Prostate	PC3, DU145
Renal	786-0, RXF-393, A498, ACHN, CAKI-1, SNC12C, TK10, U031
Unknown	ADR-RES

Since the stabilized cell lines within the NCI60 collection represent a physiological model to study gene profiles in cancer cells, with features

strongly similar to cancer tissues, we reviewed information present in public databases about the 60 cell lines and 21 normal tissues, in order to identify transformation-dependent transcriptional signatures for PKA pathway-encoding genes (Table 5).

Table 5 : ***Gene expression profiling datasets of NCI60 cell lines and normal tissues analyzed in this study.***

Reference	**Tissue of origin**	**Number of transcriptional profiles**	**GEO Number**
(Wang et al., 2006)	NCI60 cells	60	GSE5949
-	Breast	0	-
(Su et al., 2002)	CNS	2	GSE96
(Wu et al., 2007)	Colon	4	GSE6731
(Barnes et al., 2004b)	Blood	4	GSE1402
(Su et al., 2002)	Lung	2	GSE96
-	Skin	0	-
(Su et al., 2002)	Ovary	3	GSE96
(Su et al., 2002)	Prostate	3	GSE96
(Su et al., 2002)	Kidney	3	GSE96

Gene expression profiles retrieved from the GEO Database. Dataset A (60 profiles) is made up of the NCI60 cell lines (Wang et al., 2006). Dataset B (13 profiles) is a subset of transcriptional profiles of a diverse array of tissues, organs, and cell lines from a normal human physiological state (Su et al., 2002). Dataset C (4 profiles) encompasses the normal human adult samples derived from colonoscopic biopsy present in a database comprising samples of patients with Crohn's disease or ulcerative colitis (Wu et al., 2007). Dataset D (4 profiles) contains normal control samples present in a database containing transcriptional profiles of peripheral blood mononuclear cells (PBMC) obtained by juvenile arthritis patients and healthy controls (Barnes et al., 2004b).

We identified and gathered the transcriptional profile for 41 genes encoding proteins involved in the PKA pathway (adenylyl cyclases -ADCY-, phosphodiesterases -PDE-, A-kinase anchor proteins -AKAP-, cAMP-dependent transcriptional factors -TF-, PKA catalytic subunits -PRKAC- and PKA regulatory subunits -PRKACR-, Table 6) and compared expression

profiles of cancer cell lines with those of primary normal tissues, collected from different datasets (Table 6).

Table 6 : ***PKA related genes identified in all the datasets shown in Table 5 and used in this study.***

Probeset	Unigene	Symbol	Description
33353_at	Hs.192215	ADCY1	Adenylate cyclase 1 (brain)
34686_at	Hs.481545	ADCY2	Adenylate cyclase 2 (brain)
33134_at	Hs.467898	ADCY3	Adenylate cyclase 3
39383_at	Hs.525401	ADCY6	Adenylate cyclase 6
40585_at	Hs.513578	ADCY7	Adenylate cyclase 7
36246_at	Hs.414631	ADCY8	Adenylate cyclase 8 (brain)
33800_at	Hs.391860	ADCY9	Adenylate cyclase 9
37698_at	Hs.463506	AKAP1	A kinase (PRKA) anchor protein 1
36633_at	Hs.462457	AKAP10	A kinase (PRKA) anchor protein 10
34657_at	Hs.105105	AKAP11	A kinase (PRKA) anchor protein 11
37680_at	Hs.371240	AKAP12	A kinase (PRKA) anchor protein (gravin) 12
554_at	Hs.459211	AKAP13	A kinase (PRKA) anchor protein 13
41075_at	Hs.98397	AKAP3	A kinase (PRKA) anchor protein 3
37087_at	Hs.97633	AKAP4	A kinase (PRKA) anchor protein 4
32421_at	Hs.532489	AKAP5	A kinase (PRKA) anchor protein 5
40747_at	Hs.509083	AKAP6	A kinase (PRKA) anchor protein 6
41703_r_at	Hs.486483	AKAP7	A kinase (PRKA) anchor protein 7
35138_at	Hs.199029	AKAP8	A kinase (PRKA) anchor protein 8
37886_at	Hs.399800	AKAP8L	A kinase (PRKA) anchor protein 8-like
36506_at	Hs.527348	AKAP9	A kinase (PRKA) anchor protein (yotiao) 9
36297_at	Hs.435267	ATF1	Activating transcription factor 1

37535_at	Hs.516646	CREB1	cAMP responsive element binding protein 1
32066_g_at	Hs.200250	CREM	cAMP responsive element modulator
35522_at	Hs.487129	PDE10A	Phosphodiesterase 10A
36311_at	Hs.416061	PDE1A	Phosphodiesterase 1A, calmodulin-dependent
38921_at	Hs.530871	PDE1B	Phosphodiesterase 1B, calmodulin-dependent
32418_at	Hs.487897	PDE1C	Phosphodiesterase 1C, calmodulin-dependent
666_at	Hs.89901	PDE4A	Phosphodiesterase 4A, cAMP-specific
33705_at	Hs.198072	PDE4B	Phosphodiesterase 4B, cAMP-specific
38860_at	Hs.437211	PDE4C	Phosphodiesterase 4C, cAMP-specific
38526_at	Hs.117545	PDE4D	Phosphodiesterase 4D, cAMP-specific
37676_at	Hs.9333	PDE8A	Phosphodiesterase 8A
37249_at	Hs.78106	PDE8B	Phosphodiesterase 8B
33709_at	Hs.473927	PDE9A	Phosphodiesterase 9A
438_at	Hs.194350	PRKACA	Protein kinase, cAMP-dependent, catalytic, alpha
36215_at	Hs.487325	PRKACB	Protein kinase, cAMP-dependent, catalytic, beta
36359_at	Hs.158029	PRKACG	Protein kinase, cAMP-dependent, catalytic, gamma
226_at	Hs.280342	PRKAR1A	Protein kinase, cAMP-dependent, regulatory, type I, alpha
1091_at	Hs.550753	PRKAR1B	Protein kinase, cAMP-dependent, regulatory, type I, beta
116_at	Hs.517841	PRKAR2A	Protein kinase, cAMP-dependent, regulatory, type II, alpha
37221_at	Hs.433068	PRKAR2B	Protein kinase, cAMP-dependent, regulatory, type II, beta

To identify differences between normal and cancer samples, we performed an ANOVA analysis on the entire data set. As shown in Figure 17,

distributions of expression values of genes encoding proteins of the cAMP/PKA pathway were statistically different between normal and transformed cells (p-value <0.0001), indicating that in transformed cells the PKA pathway-related genes are differentially expressed as compared to normal cells. Namely, the box plot indicates that, overall, the distribution of expression of values of transformed cells is shifted towards lower expression values. Dispersion of the distribution in transformed cells is much reduced compared to that observed in normal tissues, as if transformation events superimpose a negative regulation that largely abrogates tissue-specific regulation (i.e., the major factor responsible for dispersion of expression in normal tissues, see next paragraph).

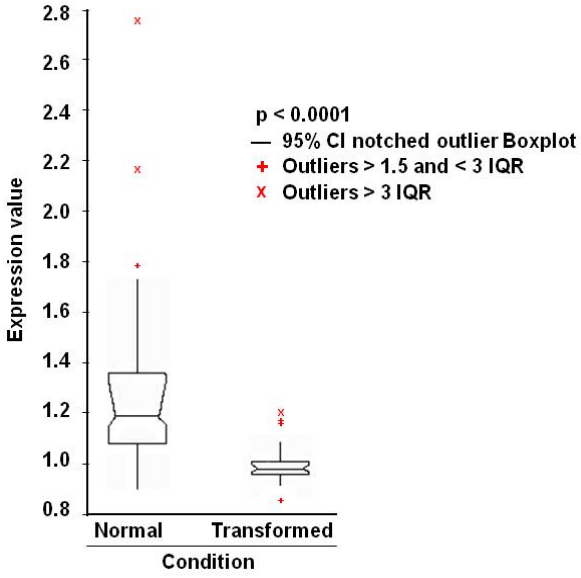

Figure 17: ***Statistical analysis of the 41 PKA pathway-encoding genes expression in normal and transformed samples****. 81 transcriptional profiles from normal tissues and from the NCI60 cancer cell line collection- were recovered from the GEO database. After normalization (see Methods), the expression values of 41 PKA pathway-encoding genes were used to perform an ANOVA analysis (p-value 0.0001) to evaluate the statistical significance of the differences between normal and transformed samples. IQR: Inter-quartile Range. Outliers are also shown.*

The same data-set was then analyzed through unsupervised hierarchical clustering (as implemented in the GeneSpring platform) that organizes genes according to the similarity or dissimilarity in expression profile, placing the cases with similar expression profiles together as neighbouring columns in the dendrogram (Figure 18).

Six different classes corresponding to the main arms of the dendrogram derived from clustering according to Tissue and cell lines (classes **I** to **III** correspond to the left main branch of dendrogram, **IV** to **VI** to the right branch) were identified. Each cell line is color-coded at the bottom according to its condition (i.e., normal, blue, or transformed, red) or the tissue of origin. Notably, classes **II** and **V** contain only transformed cells, while only one transformed cell line clusters in class **VI**. In most cases clustering effectively separates normal and transformed cell lines of the same histological origin: for instance, normal and transformed cell lines derived from kidney cluster to class **I** and **III**, hemopoietic normal and transformed cell lines to **IV** and **II**, colon cancer cells are in class **II** while normal colon in class **IV**, respectively (Table 7). Class **I** and **IV** contained cancer lines of several histological origin, while class II was enriched for cancer cells from colon and blood, class **III** for ovary and kidney and class **V** for lung, respectively (Table 7).

These results indicate that regulation of the PKA pathway is tissue-dependent, in keeping with the pleiotropic and tissue-specific phenotypes regulated by intracellular cAMP. They also suggest that transformation transcriptionally remodulates the PKA pathway, so that in most cases expression profiling of genes encoding proteins of the cAMP-PKA pathway is quite different in cancer cells as compared to their normal counterparts.

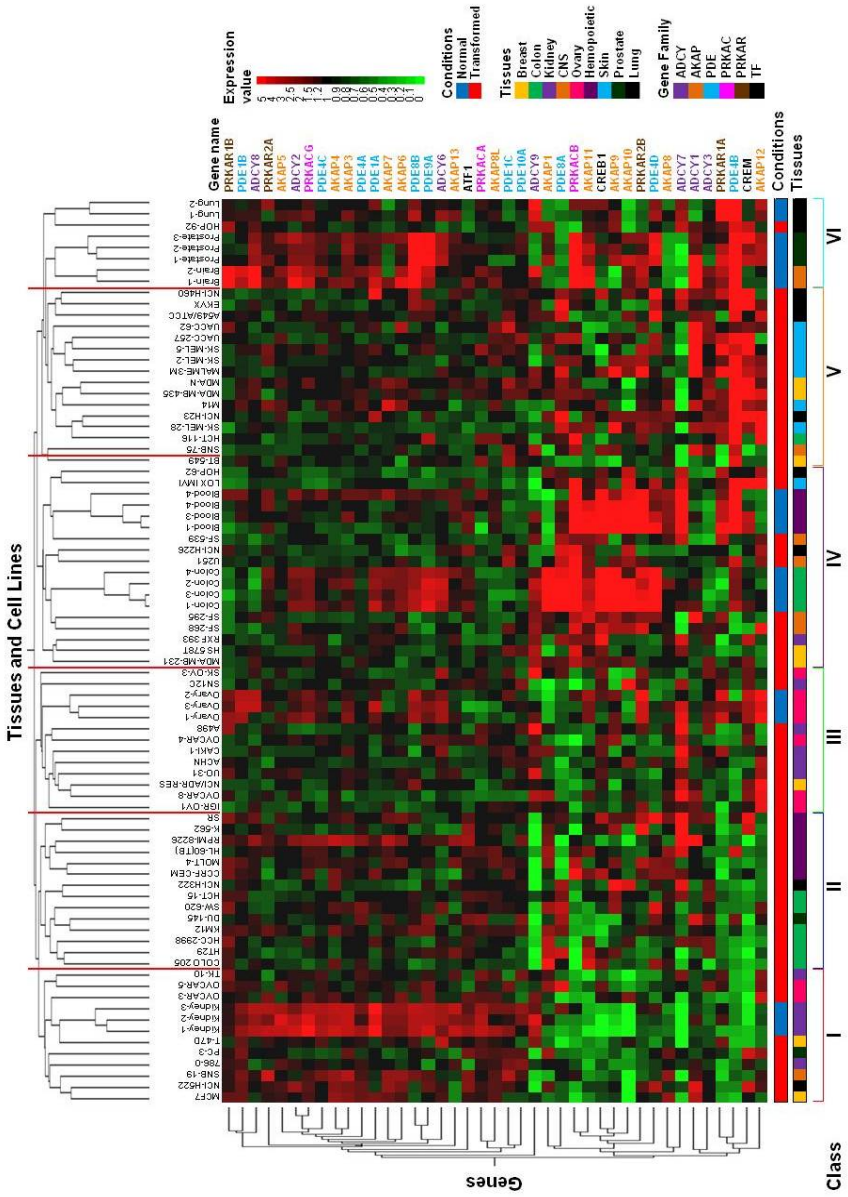

Figure 18: ***Hierarchical clustering of the 41 PKA pathway-encoding genes analyzed in this work.*** *Two-way (gene, column and cell line, row) hierarchical clustering (see Methods) of the same profiles analyzed in Figure 17. Normalized expression is colour-coded from green (poor expression) to red (strong expression). The name of each*

gene is color-coded according to family to which it belongs. The 6 main classes described in the text (red lines on the top of the dendrogram and roman number bottom of the dendrogram) are shown. The distance function is based on Pearson correlation and complete linkage clustering. Legends for expression, condition, gene family and tissue of origin are shown on the right of the dendrogram.

Interestingly in class **IV**, which comprises all the colon and hemopoietic normal samples, we observe strong expression of few genes (AKAP9-11; PDE4D; PRKCB and PRKAR2B; CREB1- colon sample- and AKAP9-11; PDE4B and PDE8A; PRKCB and PRKAR1A; CREB1 and CREM- hemopoietic sample-) as compared to their transformed counterparts, in which the same genes appeared poorly expressed (class II). In human colon carcinoma cells it has been reported that PRKAR2B over expression suppresses neoplastic cell growth (Nesterova et al., 1996), consistently with the notion that abnormal expression of isoforms of PKA regulatory subunits may be involved in neoplastic transformation. Moreover in several models of hemopoietic malignancies, it has been shown that induction of cAMP/PKA pathway stimulates leukemia cell differentiation (event associated to the relapse of the disease) or lymphoma cells apoptosis (Guillemin et al., 2002; Lerner et al., 2000).

Table 7: ***Correlation between PKA related gene patterns and tissues.***

Class	Normal	Transformed
I	Kidney 3/3	Breast 2/8 Lung 1/9 CNS 1/6 Kidney 2/8 Prostate 1/2 Ovary 2/6
II		Colon 6/7 Prostate 1/2 Lung 1/9 Hemopoietic 6/7
III	Ovary 3/3	Ovary 4/6 Breast 1/8 Kidney 5/8
IV	Colon 4/4 Hemopoietic 4/4	Breast 2/8 Kidney 1/8 CNS 4/6 Lung 2/9 Skin 1/8
V		Breast 3/8 CNS 1/6 Colon 1/7 Skin 7/8 Lung 4/9
VI	CNS 2/2 Prostate 3/3 Lung 2/2	 Lung 1/9

Six different classes, corresponding to the main arms of the dendrogram derived from clustering according to "Tissue and cell lines", were identified. Each cell line is color-coded according to its tissue of origin. The number on the right of each tissue represents the number of samples belonging to a class as compared to the total sample analyzed.

Analysis of mutational status of the NCI60 cell lines and correlation with tissue-specific PKA pathway gene regulation

In the previous paragraph we have shown that a different and a tissue-specific pattern of expression of the PKA pathway encoding genes between normal and transformed samples does exist. Moreover, we observed that a

similar pattern is common to different tissues, both in normal and transformed samples. While in normal tissues such a finding may be justified by a common histological origin or by the PKA pathway regulating a common intracellular process (i.e. differentiation, metabolism), in transformed samples, in which the correct regulation of the PKA pathway is lost, such similar gene regulation can suggest a transformation or a mutation-dependent gene regulation.

For this aim, we determined the mutation status of the NCI-60 panel of human cancer cell lines, identified the pathway in which such mutations were involved and correlated the mutation status and pathway altered in the transformed cells with transcriptional profiling data. The 60 cell lines were sorted according to mutational status, using the information provided by Catalogue Of Somatic Mutations In Cancer (http://www.sanger.ac.uk/genetics/CGP/cosmic/), and divided into four groups based on the carried mutation as follows (Table 8):

1) Cell lines carrying mutations able to interfere with the Ras pathway (i.e., mutations in genes encoding Ras, B-Raf, ERBB2, PDGFRA, referred to as Ras), 29 cell lines;

2) Cell lines carrying mutations able to interfere with PI3K-Akt pathway (i.e., mutations in genes encoding PI3KCA, PTEN and Lkb1, referred to as PI3K), 13 cell lines.

3) Cell lines carrying no somatic mutations interfering with the two above pathways (i.e., mutations in genes encoding CDKN2A, p53, referred to as Other Mutation), 14 cell lines;

4) Cell lines for which the presence of somatic mutations interfering with the above pathways has not been searched, referred to as Not Tested), 4 cell lines.

Table 8: ***NCI60 cell lines with predicted active pathways by mutational analysis.***

Tumor Type	Ras	PI3K	Other Mutation	Not Tested
Breast	HS578T, MDA-MB231, MD-MB435	MCF7, T47,	-	MDA-N, BT549
CNS	-	SF295, SF359, SNB19, U251	SF268, SNB75	-
Colon	HCC2998, HCT116, HCT15, SW620, COLO205, HT29	KM12	-	-
Leukaemia	CCRF-CEM, RPMI-8226, HL60, MOLT4, K562	-	-	SR
Melanoma	SK-MEL2, LOXIM-VI, M14, MALME-3M, SK-MEL28, SK-MEL5, UACC257, UACC62	-	-	-
Lung	A549, HOP62, NCI-H23, NCI-H460	-	EKVX, NCI-H226, NCI-H322, NCI-H522	HOP92
Ovarian	OVCAR5, OVCAR8	SKOW3, IGROV1	OVCAR3, OVCAR4	-
Prostate	-	PC3, DU145	-	-
Renal	-	786-0, RXF-393	A498, ACHN, CAKI-1, SNC12C, TK10, U031	-
Unknown	ADR-RES	-	-	-

The 60 cell lines reorganized in 4 categories, as described in the text, on the basis of the most representative mutation of each cell line: Cell lines carrying mutations able to interfere with Ras-Raf-MAPK pathway (i.e., mutations in genes encoding Ras, B-Raf, ERBB2, PDGFRA, referred to as Ras); Cell lines carrying mutations able to interfere with PI3K-Akt pathway (i.e., mutations in genes encoding PI3KCA, PTEN and Lkb1, referred to as PI3K) Cell

lines carrying no somatic mutations interfering with the two above pathways (i.e., mutations in genes encoding for CDKN2A, p53, referred to as Other Mutation); Cell lines for which the presence of somatic mutations interfering with the two above pathways has not been searched, referred to as Not Tested.

To assess overall data quality and visualize relations and differences between the aforementioned transformed and normal samples, we applied dimensional reduction through principal component analysis (PCA). A three-dimensional PCA plot of all expression data (accounting for 91% of variance) is shown in Figure 19A. PC1 (x-axis) effectively separates the normal group from the four groups of transformed cells. PC2 (y-axis) effectively separates the Ras group from the others, while PC3 (z-axis) best separates the Other Mutation group from the others. Overall, the Ras group appeared to segregate the most from the other groups.

In Figure 19B, the 41 genes encoding proteins involved in the PKA pathway were sorted according to their relative level of expression and color-coded in the graph according to expression: strong (red, value >1), average (black, value=1) and low (green, level <1). These three series were crossed with the groups described above, namely Normal, Ras, PI3K, Other Mutation and Not Tested. In Normal tissues, expression of 83% of the genes was classified as Strong, a value 2-3 fold higher than those observed in the different transformed groups (27-41%). Overall, in the transformed groups, expression of most PKA pathway-encoding genes was classified as Average or Low, with the exception of the Ras group, in which only one gene was scored as low.

Expression of PKA pathway-encoding genes was further classified as follows (Figure 19C): genes with similar level of expression between normal and at least one transformed group (blue color), genes whose expression level is different between the normal and transformed groups (yellow color) and genes with similar expression level among the different transformed

groups (grey color). Such a classification allowed us to pinpoint genes, such as ADCY2 and AKAP13 whose expression is strong only in the Normal group. More interestingly, expression of a few genes, such as ADCY3 and AKAP8 was strong only in members of the transformed groups, despite overall reduction in expression of the PKA pathway-encoding genes observed in transformed samples.

These results were further confirmed by pair-wise ANOVA analyses (Figure 19D), in which the distribution of expression values of genes encoding proteins of the cAMP/PKA pathway were found to be statistically different between normal and each group of transformed cells (p-value between 0.0001 and 0.0003). Notably, the difference in distribution between the Ras group and the PI3K and Other Mutation groups was also statistically significant, unlike the difference with the Not Tested group. This suggests that cells in this group may be biased for mutations within genes encoding proteins of the Ras pathway.

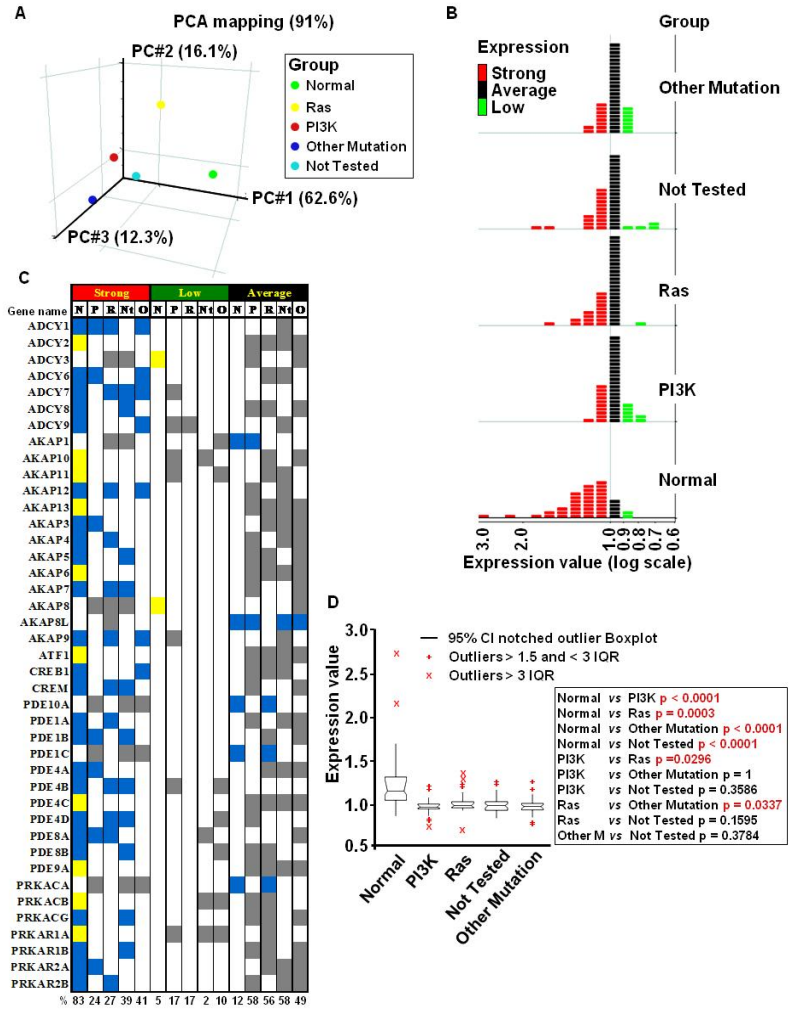

Figure 19: ***Identification of differentially regulated genes in normal and transformed samples. A)*** *Samples were sorted in five groups according to mutational activation: green, normal; yellow, Ras; red, PI3K; blue, Other Mutation; cyan, Not Tested. Principal Component Analysis (PCA) performed on 41 PKA pathway-encoding genes for normal samples and the four classes of mutation-dependent samples. Each sphere represents the comparative averaging of the 41 genes for each pathway identified by mutational analysis.* ***B)*** *For each of the 5 groups described in (A), the 41 PKA–encoding genes were clustered, relative to their level of expression, in three subgroups: Strong (>1, red), Average (=1, black) and Low (<1, green).* ***C)*** *Gene list according to expression level and mutational group of the three subgroups previously indicated, divided for each sample. Color-coding is as follows: blue, common between normal and at least one transformed sample;*

yellow, specific for normal samples; grey, specific for transformed samples. Percentage of regulated genes for each subgroup is shown at the bottom. ***D)*** *ANOVA analysis to evaluate the statistical significance of the differences between the five classes of samples described in (A). The right inset shows p-value of the pair-wise comparisons. Statistically significant differences are indicated in red. IQR: Inter-quartile Range. Outliers are also shown.*

To reveal gene expression changes relate to mutation status of the 60 cell lines, and better interpret the results of PCA and ANOVA, a hierarchical clustering was performed. The resulting dendrogram is shown in Figure 20, in which each cell line is color-coded at the bottom according to its tissue of origin -row labeled tissue-, mutated gene -row labeled mutation-, inferred pathway activated by mutation -row labeled pathway-. A robust association between the transcriptional profiles and mutations in the Ras pathway was observed (indicated as Ras, red color). Two cell lines of the Not Tested group were interdispersed within the Ras group, indicating that these two lines are most likely responsible for the lack of statistical difference between the Ras and the Not Tested group (see above). Comparison of the Tissue and Pathway categories indicated that within the two Ras sub-clusters, some tissue-specificity is conserved. Indeed, the left cluster, comprising a total of 18 cell lines, was characterized by 6 colon cancers and 6 leukemias of which 5 on 6 were mutated in Ras pathway. Similarly the right cluster, comprising a total of 19 cell lines, was characterized by 8 melanomas and 5 lung cancers of which 7 on 8 were mutated in Ras pathway for melanoma and 4 on 5 for lung cancer.

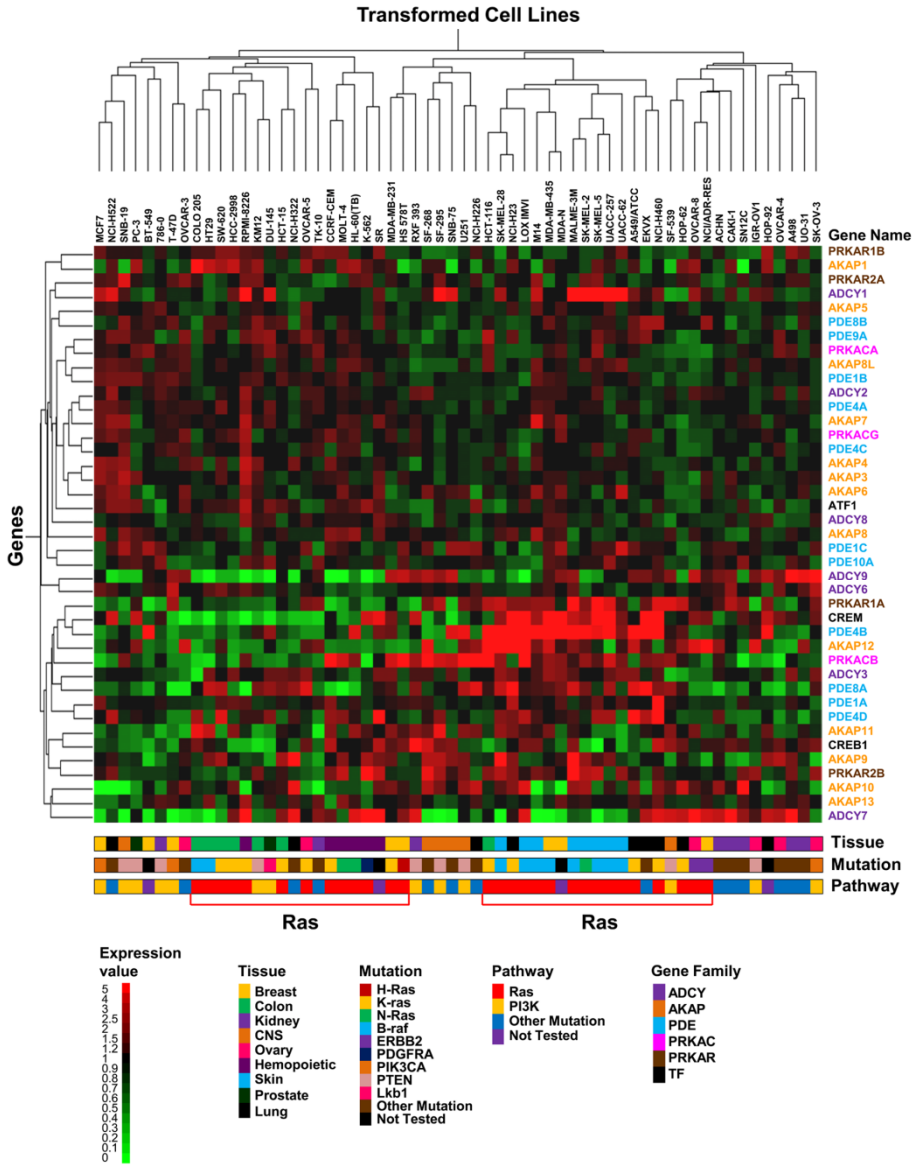

Figure 20: ***Hierarchical clustering of the 41 PKA pathway-encoding genes in transformed samples.*** *Two-way (gene, column and cell line, row) hierarchical clustering (see Methods) of the profiles from the NCI60 collection only. Normalized expression is color-coded from green (poor expression) to red (strong expression). The distance function is based on Pearson correlation and complete linkage clustering. The name of each gene is color-coded according to the family to which it belongs. Legends for expression, condition, gene family and tissue of origin are shown on the right of the dendrogram. The data have been organized on the basis of the*

tissue of origin of the cancer (Tissue), the specific oncogenic mutations identified in each cell line (Mutation), the putative altered pathway by the specific mutations (Pathway) and the gene family.

The other sub-clusters, comprising all the remaining cell lines and the other three groups of mutations and consequently of pathways, were more dispersed along the clustergram. Together, these results indicate that transformation events modulate transcriptional regulation of genes encoding proteins of the PKA pathway and that mutational activation of the Ras pathway originates a distinguishable signature, in comparison with mutational activation of the other genes studied in this report. Such a distinguishable signature is particularly noticeable in melanoma cells, in which strong expression of a gene set encoding a complete functional PKA pathway module (ADCY3; PDE4B, PDE4D and PDE8A; AKAP12; PRKAR1A and PRKAR2B; PRKACB; CREM) is observed, suggesting a deregulated cAMP signaling. Moreover, analysis of expression values for PRKAR1A and PRKAR2B genes indicated the presence in melanoma cells of a high R1/R2 ratio, that has been associated to melanocyte proliferation (Mantovani et al., 2008).

Promoter analysis: finding correlation between oncogenic pathway, transcriptional profiles and promoter regulation

Genes involved in the same pathway or transcriptionally co-regulated are likely to share similar promoter features. To test this hypothesis in our model, the 15 groups previously established (see Figure 19), containing co-regulated genes for each group, were used for promoter identification and analysis. Using a series of biocomputing procedures and statistical processes (see Methods and the Figure), we identified Transcription Factor Binding Sites (TFBSs) conserved within the promoters (operationally defined as regions spanning 500 nt upstream and 100 nt downstream from the transcription start site) of the 41 PKA pathway-encoding genes.

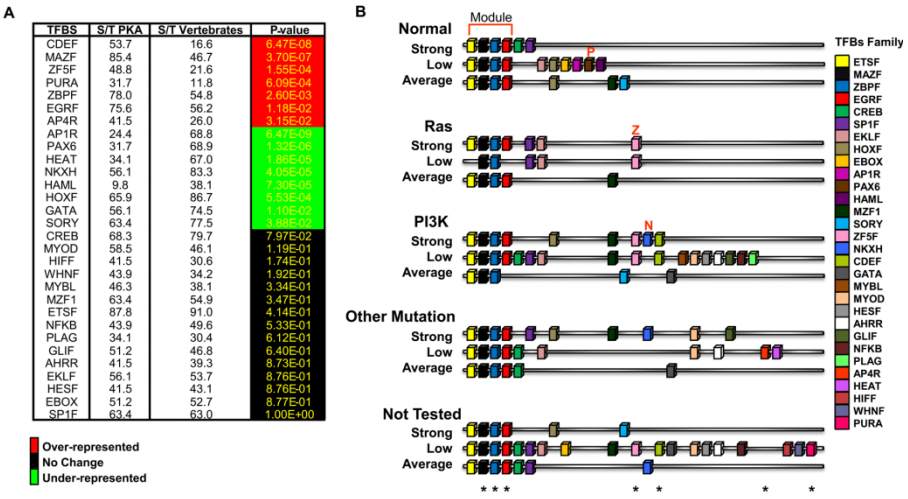

Figure 21: ***TFBS identification by using the enrichment as parameter. A)*** *The panel shows for each TFBS, recognized as relevant (present in ≥ 70% of the promoters of 41 PKA pathway-encoding genes) the percentage of promoters in our collection that contain the motif as compared to Matrix Family Library on vertebrates. This percentage has been calculated by dividing the total number of promoters containing the motif (S) by the total number of promoters (T). Color-coding scheme on the right of the panel.* ***B)*** *Schematic representation of the TFBSs (color-coded as shown on the right of the panel) identified in the promoters of the 15 subgroups described in the text and in Figure 17. Each cartoon represents the promoter structure resulting from the average of the TFBS identified in ≥ 70% of the gene promoters for each subgroup. The asterisks on the bottom of the cartoon indicate the over-represented TFBS, as scored in panel A, for all the 41 PKA pathway-encoding genes.*

Genes were sorted in the 15 groups indicated in Figure 19B and 19C, and each group separately analyzed. In this first analysis (Figure 21A), each TFBS was scored as either absent or present, regardless of the number of copies present within a given promoter. This analysis permitted the identification of 30 TFBSs enriched in the promoters of the 41 PKA pathway-encoding genes whose frequency of occurrence i.e., the ratio between the promoters that contained the specified motif (S) and the 41 promoters in our collection (T) was compared with the frequency of occurrence within vertebrate genomes (computed using the promoter Library

Matrix Family of vertebrates that comprises 260000 vertebrate promoters). Statistical analysis indicated that of these 30 TFBSs, 7 were over-represented (red color) and 9 under-represented (green color). The remaining showed the same frequency of occurrence found in the whole vertebrate genome collection.

A consensus representation for the promoter structure of each subgroup of co-regulated genes was drawn by taking into account the 30 TFBSs present in at least 70% of the genes within each subgroup (Figure 21B). Surprisingly, the vast majority of these consensus promoters (13 out of 15) showed a common module (upper part, module), comprising 4 TFBSs: ETSF, MAZF, ZBPF and EGRF, 3 of which are over-represented in our collection (over-represented motifs are indicated by an asterisk at the bottom of the figure). This strongly suggests a functional implication of these TFBSs in expression of PKA pathway-encoding genes. Other interesting features indicated by this analysis include the identification of binding sites for PAX6 (indicated by red P) and ZF5F and NKXH (indicated by red Z and N respectively) only in consensus promoters of some genes within the normal or transformed group, respectively.

Another feature that may be critical in the identification of enriched elements is the number of copies of a given TFBS within a promoter. In fact, it has been documented that the presence of multiple copies of cis-elements in promoters, particularly when clustered, makes transcriptional activation stronger (Ross et al., 2000; Wang et al., 2006).

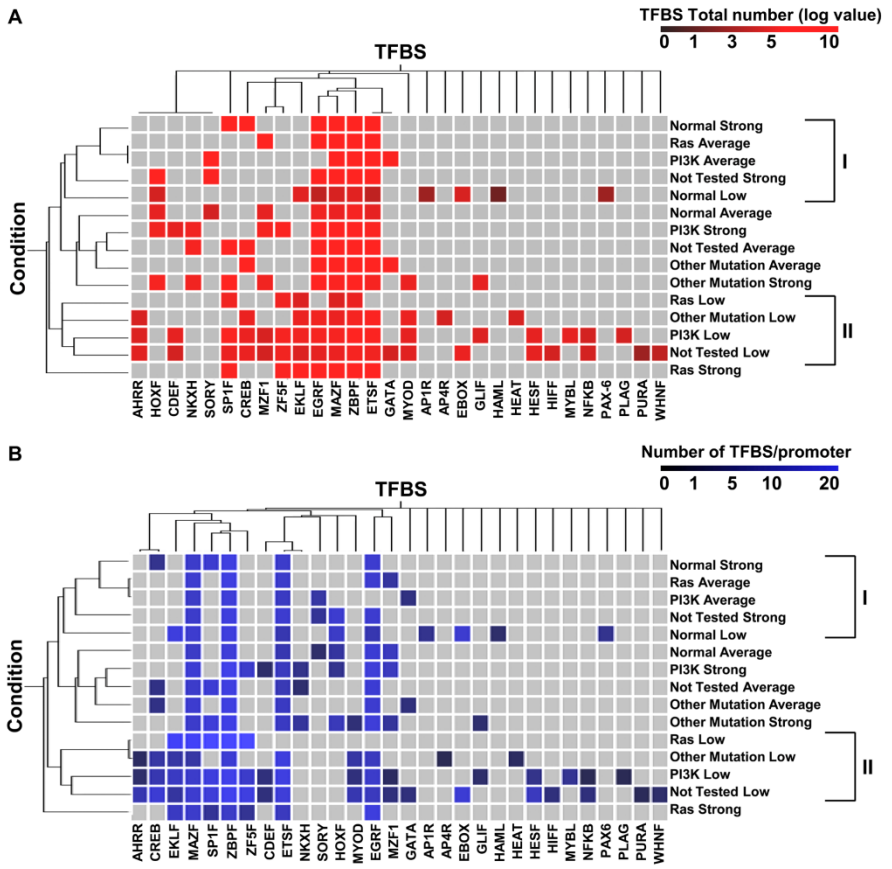

Figure 22: ***Hierarchical clustering of TFBSs present in the promoters of the 41 PKA pathway-encoding genes, according to total number and frequency.*** *Two-way (TBFS, column and expression subgroup, see Figure 16, row) hierarchical clustering of the TFBS present within the promoters of the 41 PKA pathway-encoding genes. Clustering was run according to the total number of TBFS present in each group* ***A)*** *or to the frequency, i.e. the total number of a given TBFS divided by the number of promoters* ***B)****. The color-coding scale is shown at the top of each panel. The distance function is based on Pearson correlation and complete linkage clustering. The two classes, corresponding to the main arms of the dendrogram, derived from clustering according to "Condition" are shown on the right of each dendrogram.*

For this reason, total number and frequency (number of each TFBS/promoter) of the 30 TFBSs previously identified, was scored within each of the 15 subgroups and classified by hierarchical clustering (Figure

22A and 22B, respectively). Analysis using both criteria confirmed the results reported in Figure 18: the presence in promoters of all subgroups of a TFBS module comprising ETSF, MAZF, ZBPF and EGRF. Clustering according to Regulation in Figure 20A show that all promoters of genes characterized by low expression transformed samples cluster together (class II). Promoters belonging to genes with strong expression in the Ras group cluster in a completely independent arm (lower part of the dendrogram), opposite to where cluster promoters belonging to genes with strong expression in the Normal group (class I). Additionally, clustering by frequency highlighted the specific enrichment of EKLF in genes with low expression. Clustering according to both criteria indicated that Normal samples clustered in a different way as compared to transformed samples (upper part of the dendrogram) and that the PI3K, Other Mutation and Not Tested samples were more interspersed along the dendrogram and confirmed that the Ras category showed a different promoter composition as compared to other categories, in keeping with the PCA analysis presented in Figure 16.

Data mining for PKA pathway-related gene promoters

As previously described, computational analysis of our promoter collection, permitted the identification of some TFBS that are able to characterize in a specific manner normal and transformed samples. To confirm some of our computational results, we interrogated several databases and searched in the literature for studies on promoter structure of PKA pathway-encoding genes. Experimental studies, using one or more molecular approaches including EMSA, Chromatin Immunoprecipitation and transactivation assay, have been found for 16 PKA pathway-encoding genes: PRKAR1A, PRKAR1B, PRKAR2B, PRKACA, AKAP1, AKAP8, AKAP9, AKAP10, AKAP12, ADCY8, ADCY9, PDE4B, PDE4C, PDE4D, CREB and CREM. This subset of genes was re-analyzed as described above and the obtained results were compared with literature data (Table 9). In total, 36 TFBSs have been

experimentally identified: 20 of these (i.e. 55%) have been predicted by our computational approach and for two genes alone (AKAP9 and PRKAKA), none of the experimentally identified sites was identified by the computational approach that overall identified a much higher number of sites compared to those retrieved from literature. The biological significance of the presence of the identified TFBS and of their relationship with oncogenic mutations, notably in the Ras pathway, is proposed below.

Table 9 : ***Comparison between computational data and literature data.***

GeneName	Computational data	Literature data
ADCY8	HOXF, ETSF, SORY	CREB
ADCY9	ZF5F, SP1F, ZBPF, MAZF, EKLF, EGRF, EBOX, CDEF, WHNF, AHRR, ETSF, HESF	c-Myc/EBOX
AKAP1	MYOD, EKLF, MAZF, EGRF, ETSF, HOXF, AP1R, EBOX	c-Myc/EBOX
AKAP8	SP1F, EKLF, ZBPF, EBOX, ZF5F, HIFF, MAZF, EGRF, ETSF, AHRR, CREB	CREB
AKAP9	ZBPF, EGRF, EKLF, ETSF, HOXF, SP1F, MAZF, GLIF, HOMF, NKXH, CREB	c-Myc/EBOX
AKAP10	ETSF, SP1F, HESF, EBOX, ZBPF, MAZF, MYBL, MYOD, AP4R, EGRF, CREB, MZF1	c-Myc/EBOX
AKAP12	EBOX, EGRF, SORY	c-Myc/EBOX, SRF/SRFF
CREB1	SORY, SP1F, NFKB, ETSF, ZBPF, MYOD, EKLF, AP4R, MZF1, HICF	Myc/EBOX, Sp1/SP1F, NFkB/NFKB
CREM	ZBPF, EGRF, SP1F, MAZF, EKLF, ETSF, CREB, WHNF, HESF, EBOX, MYBL, HIFF, AHRR	CREB
PDE4B	ETSF, SORY, HOXF, ZBPF, HEAT, CREB	CREB
PDE4C	ZBPF, ETSF, CREB, GLIF, MAZF	Myc/EBOX, CREB

PDE4D	SORY, HOXF, NKXH, ETSF, CREB, GATA, MYOD	CREB
PRKACA	EKLF, ZBPF, EGRF, MZF1, NKXH, ETSF, MAZF, SORY, HOXF	USF1/EBOX, USF2/EBOX
PRKAR1A	ETSF, CREB, ZBPF, EBOX, EGRF, SORY, GATA, MYBL, MYOD, HIFF, SP1F, HESF	AP1/AP1F, AP2/AP2F, Sp1/SP1F, CREB, FOXC2, FOXD, FOXD2
PRKAR1B	EKLF, MAZF, HESF, PLAG, ZBPF, EBOX, EGRF, MYOD, MZF1, AHRR, ETSF, HIFF, SP1F, AP1F	Jun/AP1F, p53/P53F, Oct-1/OCT1, Egr1/EGRF Pax1/PAX1
PRKAR2B	ZBPF, ZF5F, EGRF, EKLF, MAZF, HESF, SP1F, CREB, EBOX, ETSF, NF1F	Sp1/SP1F, NF-1/NF1F, Myc/EBOX, CEBPbeta/CEBP USF1, USF2/EBOX

The table shows the transcriptional factor families identified by computational analysis and the transcriptional factors/transcriptional family identified by analysis of literature data. The factors identified by both analysis are shown in red.

PKA type I regulatory subunit A (PRKAR1A) expression has been studied in different cellular models by analyzing its mRNA expression and by using its putative promoter region. In its promoter, binding sites for activator protein-1 and 2 (AP-1 and AP-2) and Sp1 (Solberg et al., 1997) have been identified. Moreover, a more recent work showed a direct activity of FOX family (FOXC2, D1 and D2) transcriptional factors members in the regulation of PRKAR1A expression both at transcriptional and at post-transcriptional levels (Dahle et al., 2002a; Dahle et al., 2001).

The promoter of PRKAR1B has been identified and studied in human and mouse: binding sites for Jun and p53 (human) and Oct-1, Egr1 and Pax1 (mouse) have been found. These binding sites have been experimentally verified by Electrophoretic Mobility Shift Assay, functional analysis and Northern blot (Clegg et al., 1996; Clegg et al., 1994).

PRKAR2B promoter has been studied in particular in Sertoli cells (human). Some reports identified binding sites for Sp1, NF-1, Myc, C/EBPbeta, able to induce the PRKAR2B promoter, USF1 and USF2. Interestingly, overexpression of USF2, but not USF1, led to inhibition of both cAMP- and C/EBPbeta-mediated induction of PRKAR2B (Dahle et al., 2002b; Knutsen et al., 1997; Singh et al., 1994).

The promoter of Protein kinase, cAMP-dependent, catalytic, alpha (PRKACA) has been identified both in humans and mouse, but little information has been produced for human promoter. Indeed, one paper describes the presence of binding sites for USF1 and USF2 transcription factors (Barradeau et al., 2001).

AKAP1, AKAP9 and AKAP10 promoters contain binding sites for c-Myc as shown by computational analysis and ChIP experiments in several human cell lines (Fernandez et al., 2003; Li et al., 2003). Moreover, a single study indicates the presence in the promoter of AKAP12 of binding sites for Serum Response Factor transcriptional factors (Streb and Miano, 2005) and more recently for Myc.

ADCY9 promoter contains binding sites for c-Myc as shown by an experimental approach (Mao et al., 2003).

Several promoters of genes encoding phosphodiesterase proteins have been isolated and to some extent studied. All the studies have been performed on sequences of human promoters and in particular the PDE4B, PDE4C (both present in our collection of PKA pathway related genes), and PDE5A, PDE6A, PDE6B and PDE7A promoters (not present in our gene list) have been better characterized. In the PDE4B promoter, binding sites for CREB have been found (D'Sa et al., 2002). In PDE4C promoter, binding sites for Myc have been found (Li et al., 2003). In the PDE5A promoter, binding sites for Jun and AP-2 have been found (Lin et al., 2001a; Lin et al., 2001b); in PDE6A and PDE6B promoters, binding sites for Sp1 (Mohamed et al., 1998)

and Sp4 (Lerner et al., 2001; Lerner et al., 2002) respectively and in PDE7A promoter, Ets2 and NFkB1 binding sites (Torras-Llort and Azorin, 2003).

The cyclic AMP response element (CRE)-binding protein CREB promoter has been identified in human, mouse and rat. Analysis done on human promoter, experimentally confirmed, identified binding site for c-Myc (Delfino and Walker, 1999) and Sp1 (Shell et al., 2002). Further information about such promoter has been produced in mouse and rat cells which allowed the identification of binding site for NfkB (Delfino and Walker, 1999).

An important regulative mechanism of the PKA pathway is feedback control. Indeed as well as the cAMP produced by Adenylyl Cyclases, activate PKA kinase activity, PKA is able to inhibit the pathway, activating by phosphorylation the Phosphodiesterases, which ultimately induce hydrolysis of cAMP switching off the pathway. Moreover a huge amount of data has been published regarding the ability of PKA to activate specific transcription factors by phosphorylation: cyclic AMP response element (CRE)-binding protein CREB, the cAMP response element modulator (CREM), the activating transcription factor 1 (ATF-1) and a repressor, ICER (inducible cAMP early repressor) (Mayr and Montminy, 2001) that, to a certain extent, has been shown to regulate PKA pathway-related genes transcription. Some of the promoters, already discussed above, have been shown to have CRE binding sites. Moreover, two interesting recent publications, have identified and characterized in different cellular contexts and by several approaches, through a genome-wide approach, target genes that are regulated by CREB (Impey et al., 2004). The authors have identified and proved by ChiP analysis (PRKAR1A, PDE7B) the presence of CRE site in PRKAR1A, in PDE7B, AKAP8, PDE4C and ADCY8. In the latter case they did not observe binding by Chip analysis, but another report has shown that its activation is mediated specifically via the canonical CRE site (Chao

et al., 2002). Binding sites for CREB1 have been found in PDE7A (Torras-Llort and Azorin, 2003), PDE4D (Wang et al., 2003), CREM (Walker and Habener, 1996) and experimentally confirmed. Moreover analysis of the promoter of CREB gene showed the presence of several CRE binding sites (Meyer et al., 1993).

Most of AP-1 (i.e. Jun), AP-2 and Sp1 transcription factors are involved in growth-related signal transduction pathways, among which Ras is a main actor, and their over-expression can have positive or negative effects on proliferation (Black et al., 2001; Maurer et al., 2007). Indeed Sp family has been shown to be regulated by post-translational mechanisms by Ras pathway (Pore et al., 2004) as well as Ets1 and Ets2 (Foulds et al., 2004) and NFkB (Finco and Baldwin, 1993).

Egr-1 is an early responsive gene linked to mitogenic stimulation directly regulated by MAPK pathway (Wong et al., 2002). Moreover for Myc (Sears, 2004), C/EBPbeta (Mo et al., 2004) and NF-1 (Nebl et al., 1994) a large amount of data about their correlation with Ras pathway has been reported. Each of these transcriptional factors has been associated with several cellular responses (proliferation, survival, apoptosis) and transformation as is the case of the PKA pathway as well. Therefore it is possible that mitogenic signal through Ras and the regulation of such transcription factors, modulates the expression of PKA pathway related genes.

An important role, in the activation of the CREB family transcription factors, is played by stimuli which are able to induce their phosphorylation and consequently their activation. In fact as reviewed in (Meyer et al., 1993) not only the protein kinase A is involved in this function but also several growth factors (NGF, FGF, IGF-I, PDGF, EGF), survival signals and hypoxia that often activate the Ras pathway, pointing to an essential role of the latter pathway also in gene transcriptional regulation of PKA pathway-encoding genes by transcription factors of the CREB family.

5.2.2 Discussion

By using a generalized workflow for data recovery and integration that combines accurate analysis of recovered data from several databases with the application of different statistical tests we have been able to correlate strong transcriptional repression of genes encoding proteins of the cAMP/PKA pathway in transformed samples of different genetic origin (i.e., bearing mutations in different pathways). This finding prompted us to compute consensus promoters, whose composition was specifically enriched for different transcription factor binding sites (TBFS). Comparison of TFBS computationally identified in the consensus promoters with TBFS experimentally identified by a variety of techniques, shows a good agreement. Indeed, by lowering the stringency used in the workflow, some of the TFBS missed by higher stringency analysis (false negatives) were recovered, in keeping with the notion that intersection of different data sets and/or techniques decreases both noise and the number of hits.

The workflow we have followed is summarized in Figure 23 and detailed in Methods section. As the number of sites hosting curated transcriptional profiles increases, more and more data to be used as starting point become available. We used the GEO database to recover data from the NCI60 cell collection (cancer samples) and matching normal tissues and to which specific statistical tests (i.e. ANOVA, Hierarchical clustering) were applied. By using the COSMIC database, which gives information about the mutational status of the NCI60 collection, we could sort the NCI60 cell lines in 4 subgroups with mutational activation of genes encoding components of the Ras pathway, of the PI3K pathway, of other pathways or for which no information was available. Such a sorting allowed us to uncover an hitherto unrecognized oncogene-dependent pattern of regulation of 41 genes encoding components of the cAMP/PKA pathway (Figure 23B and 23C). The transcriptional profiles for transformed cells within one of the identified

subgroups may then be used as a new query to GEO database (green arrow), in order to correlate and confirm, i.e. in cancer tissues, the oncogene-dependent pattern identified.

Deregulation of transcriptional programs, such as that identified for PKA pathway-encoding genes, may be considered a direct consequence of a deregulated activity of transcription factors. The TRANSFAC database was used with a high stringency threshold, to identify the regulatory sequence in co-regulated genes with high confidence, improving the deduced linkages between transcription factors and the regulated genes. Using this approach, we demonstrated that in all PKA encoding genes TFBSs for ETS, MAZ, ZBP and EGR transcription factors are present (Figure 23D) and that specific subsets of TFBS are present in the normal and transformed samples. The number of TFBS identified by computational analysis was higher than those that could be retrieved from literature as experimentally determined. This observation was to some extent expected because of limited literature reference availability, complexity to retrieve data, difficulty to analyze data from several origins, and the lack of powerful data analysis and integration tools. Under these less-than-ideal conditions, a dedicated tool such as the TFBS database, can be extremely powerful, allowing predictions that are amenable to experimental verification, should this be necessary. As discussed above, most of the false-negatives that failed to be detected by our computational approaches could be recovered by appropriately lowering the stringency of analysis.

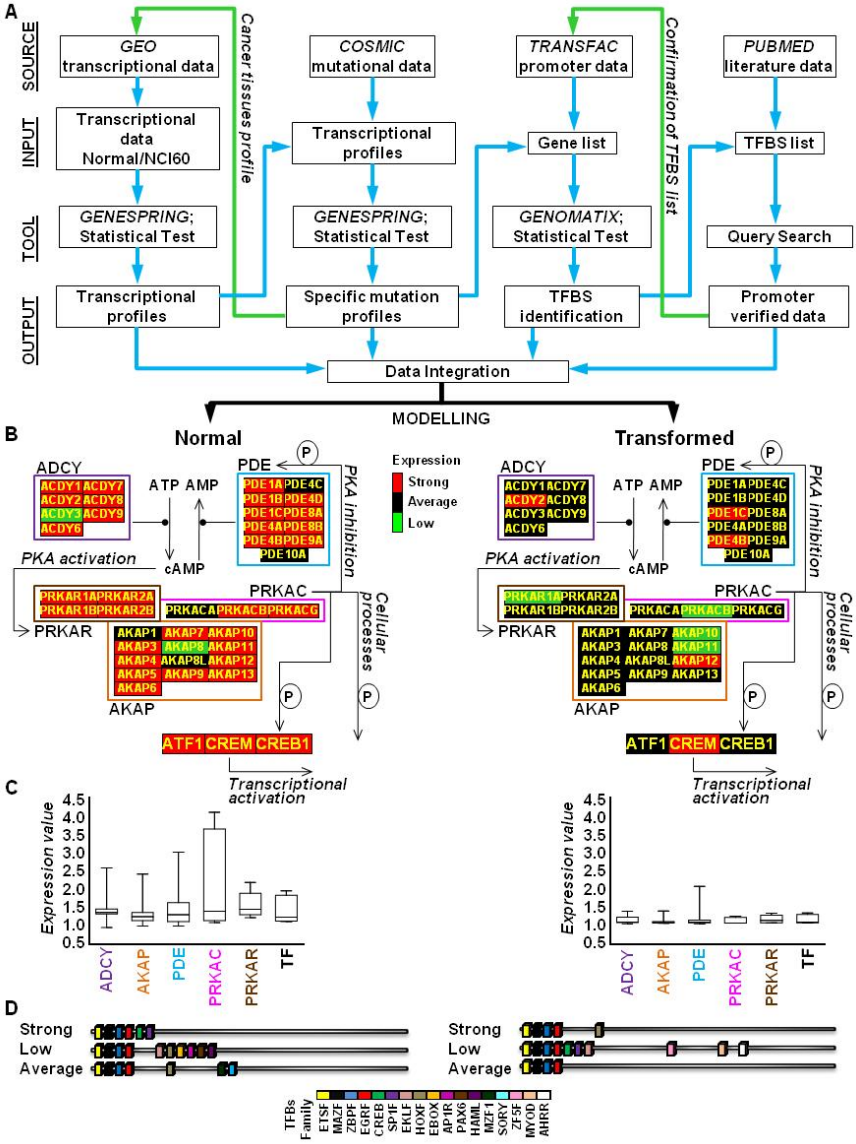

Figure 23: ***Flowchart of our web-based and statistical strategy used to elucidate the relation between PKA encoding genes transcriptional profiles and oncogenic mutations. A)*** *Flow chart of our web-based and statistical strategy with indication of some of the databases (Source) used, the type of data analyzed (Input), the specific program and statistical test (Tool) used and the result obtained (Output).* ***B)*** *Graphical representation of the block diagram summarizing functional interconnections within the PKA pathway module with indication of the expression level (geometric mean) of each gene belonging to the network -Strong (red), Average (black) and Low (green)- as identified by our analysis both in normal (B, left) and*

transformed samples (B, right). ***C)*** *Boxplots of the expression of PKA pathway-encoding genes in normal (C, left) and transformed (C, right) samples, grouped for functional classes (ADCY: adenylyl cyclase; AKAP: A-kinase anchor protein; PDE: phosphodiesterase; PRKACR: PKA regulatory subunit; PRKAC: PKA catalytic subunit). The represented value is the median.* ***D)*** *Schematic representation of the TFBSs (color-coded) identified in the promoters of PKA pathway-encoding genes of normal and transformed samples. Each cartoon represents the promoter structure resulting from the merge of the TFBS identified in ≥ 70% of the gene promoters of all normal samples and transformed samples.*

In Figure 23B transcriptional expression of PKA pathway encoding genes is color-mapped (geometric mean, Strong expression, red, Average expression, black and Low expression green) on a block diagram summarizing functional interconnections within the PKA pathway module. A general and balanced co-regulation of both positive and negative regulators of the cAMP/PKA pathway is apparent in both normal and transformed samples. Notably, in normal cells variability in expression is maximal for genes encoding the catalytic subunit of PKA. Because of the pleiotropic role of the PKA pathway (including stimulation of growth and differentiation in many cell types, such as somatotrophs, thyrocytes, melanocytes, ovarian follicular granulosa cells, keratinocytes, nervous, muscle and blood cells and adipocyte and the important role of such pathway in the regulation of the function of tissues as kidney, ovary, brain, and prostate), strong expression in normal tissues is expected (Cho-Chung, 1990; Mei et al., 2002). It should also be remembered that cross-talk between the PKA pathway and oncogene-mediated pathways can also take place at post-transcriptional levels. For example, several Authors reported the ability of oncogenic and viral Ras proteins to either stimulate (Franks et al., 1987; Spina et al., 1993) or inhibit (Beckner et al., 1985; Levitzki et al., 1986) ADCY activity in different cell lines (thyroid, epithelial, kidney, fibroblast). Moreover an involvement of MAPK or PI3K pathways in the regulation of PDE activity has been reported, suggesting that mitogenic stimulation may

positively regulate PDE4 expression directly (Liu et al., 2000), confirming our transcriptional results, or by post-translational mechanisms in which p42(MAPK) phosphorylation activity has a relevant role in their regulation (Houslay and Baillie, 2003). Another important post-transcriptional mechanism that links Ras or PI3K pathways to cAMP/PKA pathway is the positive and negative control of CREB activity by a phosphorylation (Salas et al., 2003). Moreover, it has been reported that cAMP is able to induce proliferation rather than growth inhibition, in several tumors where oncogenic activation of B-Raf has been identified (i.e., melanoma and thyroid cancer) (Dumaz and Marais, 2005). Nevertheless, the general and coordinated down-regulation of essentially all genes of the pathway in transformed cells (as compared with normal tissues) suggests that at least one PKA-mediated function needs to be reduced substantially in order to express the transformed phenotype. Although at this stage it is too early to propose specific hypotheses, it is intriguing to remember that PKA has been ascribed a role in activating mitochondrial respiration and decreasing ROS production (Raha et al., 2002), thus effectively counteracting mitochondria dysfunction that is found associated with increased glycolysis (Warburg effect, (Chiaradonna et al., 2006a; Warburg, 1956)) in many cancer cells. On the other hand, a reduction in oxidative phosphorylation that will decrease ATP supply, as substrate of adenylate cyclase, may result in a decreased cAMP production without relevant changes in the level of the enzymes (and possibly therefore of their gene expression).

6. Identification of phylogenetic conserved characteristic of cancer cells by comparison analysis between mouse and human species.

The mutations that cause cancer trigger thousands of perturbations in cell circuitry components, such as changes in gene expression levels. Animal models have proved to be useful experimental tools for identifying and characterizing such of these genetic factors. Such a comparative approach helps to identify factors that animals and man have in common and to elucidate the basic mechanism of tumorigenesis. More recently, gene expression profiling has attracted interest as a means of comparing the molecular features of tumors among different vertebrate species as dog, mouse and rat. For example, Sweet-Cordero et al. (Sweet-Cordero et al., 2005) used comparative functional genomics to assess the molecular relationship of mouse models of KRAS2-mediated lung cancer or hepatocellular carcinoma to their human correlates. These studies reveal two important facts: (i) some animal models more closely recapitulate their corresponding human tumors than do others, which determines their relative comparative value, and (ii) the comparison of human cancer expression profiles with those found in valid animal tumor models can reveal previously unrecognized gene expression signatures. The mouse remains the animal model of choice for several reasons. The principal it is that mice and humans have roughly the same number of genes, and intracellular signaling pathways are highly conserved between the two species.

Mouse models of cancer provide us with the ability to learn about tumor biology in complicated and dynamic physiological systems. Even though tumorigenesis in mice does not fully parallel that in humans, is it possible identify the signal coming from such a cross-species comparison would be analogous to the detection of conserved regulatory regions by comparative

genomics. In fact cross-species studies using genomic-based technologies have indicated the preservation of oncogene transcriptional signatures (Ellwood-Yen et al., 2003; Sweet-Cordero et al., 2005) or the synteny of tumor-associated copy number alterations (Hodgson et al., 2001; Maser et al., 2007; O'Hagan et al., 2002). Furthermore, comparison between mouse and human samples have demonstrated the conservation of somatic signature mutational events (Maser et al., 2007; O'Hagan et al., 2002), and have enabled the efficient identification of new oncogenes in human cancers (Menendez et al., 2009).

In this chapter, we discuss the use of comparative analysis to advance in the cancer biology, in particular this analysis has been performed and described in two works: *Comparative transcriptional analysis between a K-ras mouse cell model of transformation and the NCI60 human cancer cells collection* and *Promoter Scan: Algorithm to detect over-represented TFBSs in the proximal promoter regions of co-regulated or co-classified.*

In the first work, in order to identify common gene expression signatures between a set of 60 human cancer cell lines (NCI-60 cell collection) and a mouse cell model of oncogenic K-ras dependent transformation, has been carried out a comprehensive transcriptomic analysis. All the microarray data have been first normalized by using the RMA procedure and then analyzed by using advanced multivariate statistical methods, such as ANOVA, PCA, clustering and pathway analysis tools (as described elsewhere in the thesis), to identify the differentially expressed genes and the pathways to which they are associated. By focusing on the similarities between the two species, we showed that between transformed mouse and human cell models, especially that depending on K-ras for transformation, exists a strong correlation at both transcriptional and pathway levels. These findings demonstrate the power of the comparative analysis as described here as a tool to identify cancer-specific gene

signatures. Our results further show that our mouse model can be used to study the human oncogenic process.

In the work “Promoter Scan: Algorithm to detect over-represented Transcriptional Factor Binding Sites (TFBSs) in the proximal promoter regions of co-regulated or co-classified genes”, we presented a new algorithm that may give an important contribute in the important problem in understanding the complex nature of eukaryotic gene regulation. We describe a new and flexible framework for identification of common structure of regulatory elements inside the proximal promoter region of co-regulated genes. The strength of this algorithm lies in two new parameters: the conservation and the physical characteristics of the putative transcription factors, that permit a strong reduction of predicted Transcriptional Factors (TFs) without reduction of the “true” TFs.

6.1 Comparative transcriptional analysis between a K-ras mouse cell model of transformation and the NCI60 human cancer cells collection.

The more of 22000 genes in the mammalian genome, acting combinatorial within individual cells, are able to create the extraordinarily complex organismic phenotypes of mammalian body. A central goal of twenty-first-century biology is to relate the functioning of this large repertoire of genes to organismic physiology, developmental biology, and disease development. In this way, the microarrays experiments permit us to look at overall and characterize the transcriptional profiles of tens of thousands of genes simultaneously, this technologies have been widely used in biomedical (Alon et al., 1999; Golub et al., 1999; van 't Veer et al., 2002) and comparative genomic studies (Bergmann et al., 2004; Lelandais et al., 2006; Zhou and Gibson, 2004).

The widespread application of DNA microarrays to cancer research is nothing less than astounding. In the short ten years history of this versatile

technology, hundreds of large-scale experiments have been done, generating global quantitative profiles of gene expression implicated in cancer cell growth, survival, progression, metastatic invasiveness and/or therapeutic resistance (Hanash, 2004; Rhodes and Chinnaiyan, 2005; Segal et al., 2005). The nature of cancer suggests that it is a disease of chaos, a breakdown of existing biological order within the body. More specially, the disorder observed in cancer appears to derive directly from malfunctioning of the controls that are normally responsible for determining when and where cells throughout the body will multiply. In general, cancer is a heterogeneous cellular disorder caused by the deregulation of many interacting cellular pathways that converge to generate tumor formation and growth.

Known types and subtypes of cancer have been readily distinguished by their gene expression patterns, and more importantly, new molecular subtypes of cancer have been discovered that are associated with a host of tumor properties, including the propensity to metastasize and sensitivity or resistance to particular therapies. Although human cancers harbor hundreds of genetic alterations, only a subset of these alterations is likely to impact tumor initiation or maintenance. A new line of attack seeks to examine the cancer profile as a whole, often in the context of other cancer signatures or other types of genomic data. Such integrative approaches are capable of simplifying complex cancer signatures into co-ordinately regulated modules, transforming one-dimensional cancer signatures into multidimensional interaction networks and extracting regulatory mechanisms encoded in cancer gene expression.

Comparative genomics adopts the assumption that important biological processes are often conserved across related species. Based on that, scientists use animal models to infer human physiological and genetic properties (Bedell et al., 1997a; Bedell et al., 1997b; Meuwissen and Berns, 2005; Sell, 2003). Using microarray data, some theories on gene expression evolution

across genomes have been suggested. For example, Khaitovich et al (Khaitovich et al., 2004) proposed that the majority of expression divergences between species are selectively neutral and are of no functional significance. The more of these studies deviated from the idea that genes should be expressed properly to conduct their functions and that basic biological processes are often conserved between related species. Jordan et al (Jordan et al., 2005) suggested that gene expression divergence among mammalian species is subject to the effects of purifying selective constraint, and it could also be substantially influenced by positive Darwinian selection. Liao and Zhang (Liao and Zhang, 2006a; Liao and Zhang, 2006b) found that the expression profile divergence for the majority of orthologous genes between humans and mice is significantly lower than expected under neutrality and is correlated with the coding sequence divergence.

This study describes the development and application of genome-scale high-throughput methods in order to find frameworks, which can be used to investigate the expression data across genes and across platforms from human and mouse genomes. Here we describe the identification, by using several methods including multivariate statistical methods, such as ANOVA, PCA, and clustering and pathway analysis tools, of a nutshell of genes that can be considered a baseline matrix of transformation process, conserved between several type of cancer cells and between rodent and human cancer cells.

6.1.1 Results

Global analysis of human and mouse datasets

Using the Affymetrix array technologies, expression profiles analysis has been performed starting from the dataset created for human and mouse collection data, divided into 3 categories: normal tissues '**NT**', cancer '**CCL**' and immortalized '**ICL**' cell lines (see Table 10).

Table 10: ***Human and mouse genechip data collection***

	Cancer Cell Lines (CCL)		**Immortalized Cell Lines (ICL)**		**Normal Tissues (NT)**	
	N°	**GEO ID**	**N°**	**GEO ID**	**N°**	**GEO ID**
Breast (BR)	7	*	3	GSM200612 GSM200739 GSM50033	1	GSM44683
Central Nervous System (CNS)	6	*	-	-	3	GSM18921 GSM18922 GSM44690
Colon (CO)	7	*	-	-	1	GSM44680
Lung (LC)	8	*	5	GSM185872 GSM427197 GSM427198 GSM427200 GSM427201	2	GSM18949 GSM44704
Leukemia (LE)	6	*	-	-	1	GSM18868
Melanoma (ME)	8	*	2	*	1	GSM44686
Ovary (OV)	7	*	-	-	3	GSM18997 GSM18998 GSM44674
Prostate (PR)	2	*	-	-	2	GSM18958 GSM44678
Renal (RE)	8	*	-	-	1	GSM44675
Mouse Embryo Fibroblasts (MEF)	-	-	-	-	2	GSM160088 GSM160104
Mouse Model	2	+	2	+	-	-

** Data obtained from http://discover.nci.nih.gov/cellminer/home.do*
+ Data obtained from our lab.

The first part of the analysis focused on identification of global differences among NT, CCL and ICL. In particular, by integrating the results of different

statistical tests, the global behavior of these 3 groups, both in human and mouse samples, have been identified. In order to describe the variability within the human and mouse sample sets, PCA analysis, using entire sample sets of array (composed by 22283 probe sets for human and 45101 for mouse), without filtering of noise data, has been used. In particular for human database the samples have been classified according to the *Type* of mutations -*Class*- (as identified in (Balestrieri et al., 2009)) and to the Tissues of origin. A three dimensional PCA plot of all expression data (accounting for 90% of variance) is shown in Figure 24A.

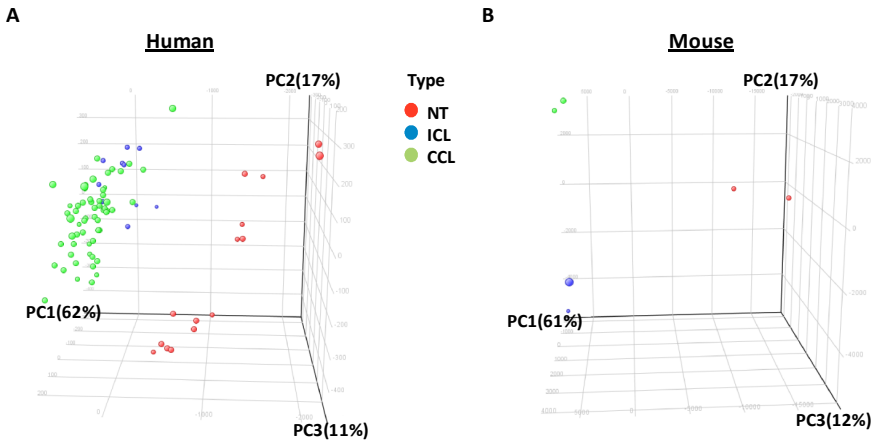

Figure 24: ***Principle Component Analysis (PCA) of human and mouse samples across the entire sample set of array. A)*** *and* ***B)*** *show the distribution on 3D space of CCL (green), NT (red) and ICL (blue), in human and mouse respectively.*

As shown in the figure, the hNT (red), the hICL (blue) and the hCCL (green) samples lay in different regions of the space and cluster separately. This result is more clearly observed in Figure 24A, where the bidimensional scatter plots between 2 components (PC1*vs*PC2, PC1*vs*PC3 and PC2*vs*PC3), show that PC1 (x-axis) separates the hNT from the hCCL in two markedly different portions of space, suggesting high differences in their gene expression values. However, the hICL samples occupy a larger portion of the available 2D space, suggesting high variability in gene expression for such

sample. PCA analysis, have been also applied to mouse data. In this case, the sample collections have been classified only by Type (MEF - mouse embryo fibroblasts- that represent normal tissue 'mNT', cancer 'mCCL' and immortalized 'mICL' cell lines). The three dimensional PCA plot of all expression data (accounting for 90% of variance), shown in Figure 24B, as well as the bidimensional scatter plots (Figure 25B), show that PC1 (x-axis) separate the mNT from the mCCL and mICL samples in two markedly different portions of space, suggesting high differences in their gene expression values. Moreover the PC2 is able to cluster mICL and mCCL in two different regions of spaces, suggesting a transcriptional rearrangement for the two samples.

The similarity of the two PCA results demonstrated the ability of mouse models to recapitulate the global transcriptional behavior of human data. In particular, our data indicate that expression patterns across human and mouse Type are conserved and hence suggesting that mouse cancer cell lines may be a helpful representation model of human tumor cells in gene expression profile studies. Our findings agree with some other comparative studies that indicated a conservation of mammalian cancer diseases (Ellwood-Yen et al., 2003; Garcia-Escudero et al., 2010; Hood et al., 2004; Klein et al., 2007; Lee et al., 2004; Miller et al., 2004; Pritchard et al., 2009; Strand et al., 2007; Sweet-Cordero et al., 2005). Conservation of patterned gene expression in the mammalian cancer models is consistent with standard assumptions of biological uniformity justifying the use of model organisms.

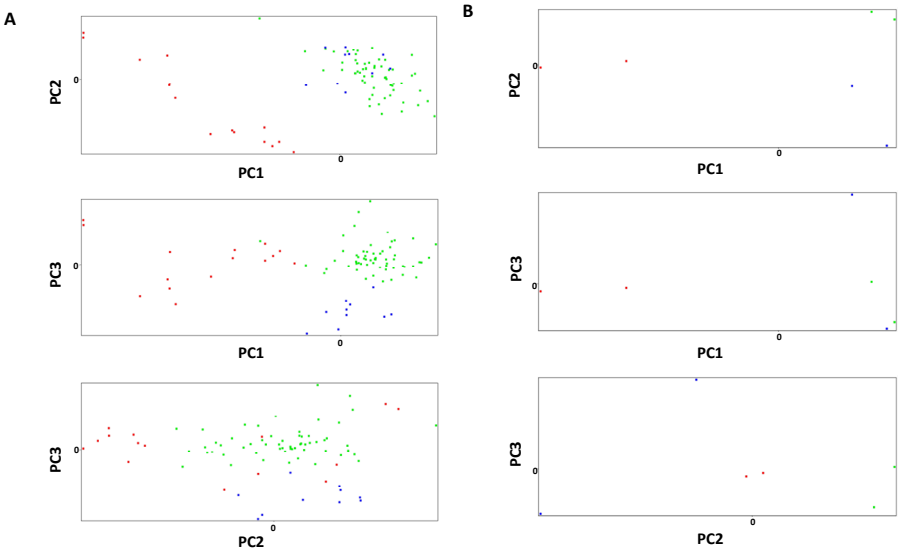

Figure 25: ***Principle Component Analysis (PCA) of human and mouse samples across the entire sample set of array****.* ***A)*** *and* ***B)*** *show the distribution on 2D space of CCL (green), NT (red) and ICL (blue), in human and mouse respectively, where the x- and y- axes show the range of variability of the two principle components (PC1 and PC2 or PC1 and PC3 or PC2 and PC3).*

The first cross-species analysis

To perform mouse-human comparison, orthologous probe sets on the two microarrays Affymetrix GeneChip, Mouse 430 2.0 (Mouse) and HGU133A (Human), have been initially found (see Methods). It was necessary to compare the expression on a gene-by-gene basis across the human and mouse arrays. This was complicated by the fact that genes are often represented by more than one probeset on each array. Genes were identified using the corresponding Affymetrix annotation and symbol on each array. Essential requirement for this procedure was that both human and mouse gene annotation and symbol had to be identical, since we consider this assumption as the only criteria to find the orthologous pairs. This procedure was performed at The NetAffx™ Analysis Center and permitted the identification of 43514 mouse probe sets (from a total of 45101 on Mouse

430 2.0 array) to one or more orthologous human probe sets on HG-U133A array (22283 probe sets). Using the same criteria, the 17331 human probe sets present in the HG-U133A array were mapped to one or more orthologous mouse probe sets. These two lists were submitted separately to flag and gene filtering. Such a procedure permitted the recognition of 10478 human and 13382 mouse probe sets respectively. Altogether the probesets from both species concurred to the generation of a new list of 21606 mouse-human orthologue probe sets (for a more extensive description, see Methods).

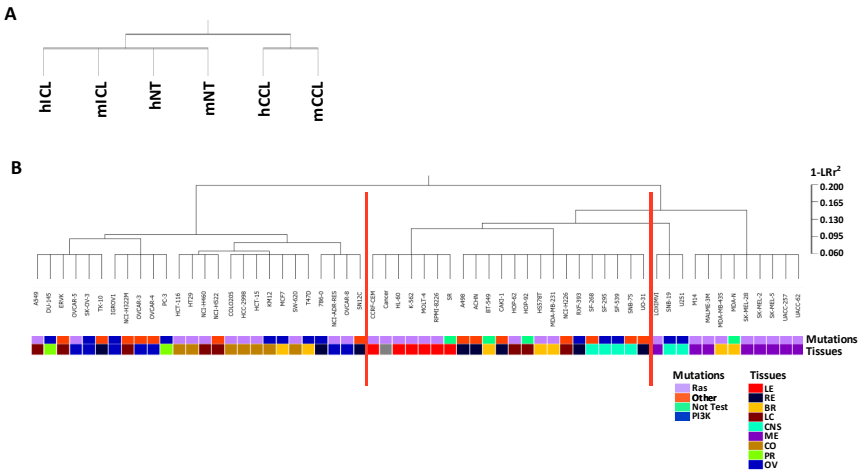

Figure 26: ***SOM and hierarchical cluster analysis of mouse and human cancer cell lines. A)*** *clustering analysis using the 21606 mouse-human orthologue probe sets from mouse and human samples classified by Type and* ***B)*** *by Mutations. Both are based on log ratios with a distance metric of (1-LRr^2), where LRr^2 is the squared Pearson correlation coefficient between the log ratios.*

Self Organizing Map (SOM) analysis and hierarchical clustering of CCL with mouse-human orthologue probe sets were performed. As shown in Figure 26A the human and mouse samples (composed of the same collection of probe sets) of the same Type (i.e., hCCL and mCCL) clustered in species-independent manner in a narrow region. In addition, hierarchical cluster of all individual samples (Figure 26B) showed that the mouse cancer model

(green) clustered with human cell lines encompassing Ras mutation (i.e. CCRF-CEM, HL-60, MDA-MB231, ACHN etc).

Cross-species analysis by altered genes and pathways.

To identify those genes that had significant differences in the level of expression between CCL-NT comparisons ANOVA analysis was used. The analysis was performed using the 13382 mouse probesets and the 10478 human probe sets separately; a p-value cut-off 0.02 and a fold change filtering ($\geq +2$ or ≤ -2). Furthermore, to minimize false-positive cases Benjamini-Hochberg multiple testing correction was applied (FDR ≤ 0.1). In order to select only one probe set for each gene in each species, the p-values obtained from the ANOVA analysis were used. In particular, we selected the probe sets with the lower p-value for both species (see Methods).

*Table 11: **ANOVA analysis.** The numbers on the table represent modulated genes in the CCL and NT comparison pairs. Up regulated genes have a fold change $\geq +2$ and down regulated have a fold change value ≤ -2.*

	Total Gene	**UP**	**DOWN**
mCCL/mNT	1905	1460	445
hCCL/hNT	1774	1386	338

Table 11 provides a summary of the significant genes differentially expressed for both human and mouse. As shown the total gene differentially expressed were similar between the two species and represent a statistical variation on expression of 14-16% of starting probesets. It is interesting to note the vast majority of differentially expressed genes are up-regulated more than down regulated. This observation can suggest that the transformation process, at least at transcriptional level, is an active mechanism. Starting from these genes (Total Gene in Table 11), obtained by ANOVA, we performed two analyses across species: the first based on pathway-specific and the second on fold change comparison.

Pathway identification by analysing the genes present in the comparisons CCL-NT (mouse and human), was performed by using GeneCodis tools. A hypergeometric test and a p-value cut-off 0.05 with FDR ≤ 0.05 were used as statistical parameters to define significantly changed pathways. This procedure predicted 98 and 78 changed pathways for mouse and human species respectively (data no show). Notably 56 pathways were found in both species, reported in Table 12, suggesting a good degree of similarity between the two models of transformation as compared to their normal counterparts. Moreover, 72% of the pathways identified in cancer mouse gene list were present also in cancer human data.

*Table 12: **KEGG pathway**. For each pathway is indicated the number of genes changed in the two species (Human and Mouse genes) and the p-value correct ≤ 0.05. Significance has been evaluated by hypergeometric distribution in both species.*

	H genes	**M genes**	**Human p-value***	**Mouse p-value***
Pathways in cancer	46	71	8.66E-07	4.16E-20
Focal adhesion	29	50	6.78E-05	4.49E-17
Huntington's disease	51	44	7.23E-20	3.31E-14
MAPK signaling pathway	30	54	3.39E-03	5.55E-14
Prostate cancer	18	28	4.22E-05	3.93E-12
Small cell lung cancer	13	26	6.10E-03	3.77E-11
Oxidative phosphorylation	37	33	1.48E-14	5.22E-11
Alzheimer's disease	40	37	2.00E-13	1.30E-09
ErbB signaling pathway	11	24	3.86E-02	2.46E-09
Endocytosis	21	38	1.28E-02	2.66E-09
Chronic myeloid leukemia	12	22	6.66E-03	3.53E-09
Cell cycle	54	28	2.53E-31	6.80E-09
Nucleotide excision repair	17	16	3.64E-09	1.15E-08
Neurotrophin signaling pathway	18	28	2.50E-03	1.28E-08
Colorectal cancer	14	22	2.39E-03	4.54E-08
Parkinson's disease	37	28	7.32E-15	5.49E-08
Non-small cell lung cancer	8	17	3.86E-02	7.34E-08
Renal cell carcinoma	10	19	2.71E-02	9.99E-08
Purine metabolism	39	30	4.79E-13	1.06E-07

Glioma	12	18	1.89E-03	1.22E-07
Melanoma	12	19	4.01E-03	1.46E-07
Peroxisome	10	20	4.41E-02	1.68E-07
Pyrimidine metabolism	27	22	1.23E-10	2.24E-07
Regulation of actin cytoskeleton	31	34	3.28E-05	6.38E-07
Axon guidance	15	25	3.15E-02	1.12E-06
Ubiquitin mediated proteolysis	25	25	4.43E-06	1.98E-06
Acute myeloid leukemia	8	15	4.43E-02	4.17E-06
Wnt signaling pathway	20	25	3.44E-03	8.86E-06
RNA polymerase	11	10	4.59E-06	9.00E-06
Homologous recombination	10	10	2.77E-05	9.00E-06
Spliceosome	59	22	2.92E-36	9.84E-06
Mismatch repair	12	9	2.17E-08	1.04E-05
DNA replication	22	11	8.61E-17	1.55E-05
Proteasome	36	12	1.04E-34	2.00E-05
Aldosterone-regulated sodium reabsorption	9	12	3.37E-03	2.00E-05
Fc gamma R-mediated phagocytosis	12	18	3.12E-02	3.01E-05
p53 signaling pathway	18	15	9.18E-07	3.18E-05
Progesterone-mediated oocyte maturation	17	17	7.60E-05	3.24E-05
Pyruvate metabolism	11	11	1.12E-04	5.21E-05
Biosynthesis of unsaturated fatty acids	6	9	6.22E-03	5.58E-05
Chemokine signaling pathway	20	26	2.93E-02	6.03E-05
Citrate cycle (TCA cycle)	10	9	4.73E-05	1.70E-04
ECM-receptor interaction	14	15	2.39E-03	2.18E-04
Long-term potentiation	10	13	2.25E-02	4.43E-04
Oocyte meiosis	27	17	3.60E-09	5.60E-04
Long-term depression	11	13	9.86E-03	5.65E-04
Amino and nucleotide sugar metabolism	7	10	3.90E-02	5.66E-04
Steroid biosynthesis	4	6	4.06E-02	7.99E-04
Gap junction	13	14	8.78E-03	1.45E-03
Bladder cancer	8	9	1.16E-02	1.66E-03
Valine, leucine and isoleucine degradation	8	9	1.37E-02	3.22E-03
Glutathione metabolism	12	9	7.13E-05	8.28E-03
RNA degradation	22	9	1.15E-11	1.30E-02
Terpenoid backbone biosynthesis	5	4	5.89E-03	1.61E-02
Arginine and proline metabolism	9	8	1.33E-02	2.47E-02
Base excision repair	12	6	3.88E-06	3.13E-02

To further detail the pathway analysis, the identification of significant pathways have been done by using up- and down-regulated genes respectively. As shown in Tables 13 and 14 several pathways were common and significantly different in both species in (CCL-NT) comparisons. In fact, in both species, we observed up regulation of pathways involved in RNA modifications and synthesis (Spliceosome, RNA degradation and RNA polymerase), protein degradation (Proteosome and Ubiquitin mediated proteolysis), cell cycle regulation (Cell cycle, DNA replication, Progesterone-mediated oocyte maturation and Oocyte meiosis), cellular metabolism (Oxidative Phosphorylation, Pyrimidine and Purine metabolism, TCA cycle, Pyruvate metabolism, Glutathione metabolism, Biosynthesis of unsaturated fatty acids, Terpenoid backbone biosynthesis, Amino sugar and Nucleotide sugar metabolism and steroid biosynthesis), DNA repair mechanisms (Nucleotide excision repair, Mismatch repair, Base excision repair, Homologous recombination and Non-homologous end-joining) and several disease-associated pathways (Huntington's disease, Parkinson's disease, Alzheimer's disease, Chronic myeloid leukemia, Small cell lung cancer and Prostate cancer). Altogether these pathways have been closely related to the onset and maintenance of transformed phenotype. In example changed metabolism has been observed in several cancer cells and tissues (Jezek et al., 2010); increased RNA and DNA synthesis (proliferative response) as well as deregulation of DNA repair mechanisms also have been associated to tumorigenesis (Herzig and Christofori, 2002; Nagaraju and Scully, 2007; Somyajit et al., 2010) It is interesting to note that between the down-regulated pathways, many are involved in processes related to cell adhesion and migration and cell to cell crosstalk (i.e., Focal adhesion, ECM-receptor interaction, Regulation of actin cytoskeleton. Gap junction, Axon guidance and Leukocyte transendothelial migration) suggesting an important role of these processes in tumorigenesis. In fact the central role of migration and invasion machineries for tumor metastasis has been showed in literature

(Cho and Klemke, 2000; Clark et al., 2007; Friedl and Wolf, 2003; Geho et al., 2005; Le Devedec et al., 2010; Pantel and Brakenhoff, 2004).

Table 13: ***Up regulated pathways modulated in both species.***

	n° genes human	**FDR human**	**n° genes mouse**	**FDR mouse**
		UP- regulated		
Spliceosome	58	4.83E-41	22	3.79E-07
Proteasome	36	1.42E-38	12	2.23E-06
Cell cycle	51	3.18E-33	17	1.45E-04
Huntington's disease	48	6.07E-22	40	9.31E-15
DNA replication	22	4.21E-19	11	1.90E-06
Parkinson's disease	37	2.29E-18	27	1.78E-09
Oxidative phosphorylation	37	4.81E-18	33	1.07E-13
Alzheimer's disease	37	1.31E-14	34	1.38E-10
RNA degradation	22	8.23E-14	8	1.01E-02
Pyrimidine metabolism	27	3.92E-13	19	6.82E-07
Purine metabolism	34	1.12E-12	22	1.47E-05
Nucleotide excision repair	17	7.89E-11	16	5.76E-10
Mismatch repair	12	1.31E-09	9	1.96E-06
Oocyte meiosis	23	1.59E-08	14	1.21E-03
Ubiquitin mediated proteolysis	25	4.33E-08	23	6.96E-07
Base excision repair	12	2.64E-07	5	4.14E-02
RNA polymerase	11	4.00E-07	9	1.09E-05
Homologous recombination	10	2.90E-06	10	1.40E-06
Citrate cycle (TCA cycle)	10	5.51E-06	9	3.29E-05
Pyruvate metabolism	10	8.82E-05	10	4.20E-05
Non-homologous end-joining	6	9.56E-05	3	3.97E-02
Glutathione metabolism	10	2.47E-04	8	7.15E-03
Progesterone-mediated oocyte maturation	14	2.70E-04	11	4.54E-03
Biosynthesis of unsaturated fatty acids	6	2.32E-03	8	7.83E-05
Terpenoid backbone biosynthesis	5	2.48E-03	4	7.78E-03
Pathways in cancer	29	4.44E-03	45	5.88E-10
Chronic myeloid leukemia	10	1.28E-02	15	7.93E-06
Amino sugar and nucleotide sugar metabolism	7	1.90E-02	9	4.58E-04
Small cell lung cancer	10	2.50E-02	17	1.69E-06
Steroid biosynthesis	4	2.74E-02	6	2.52E-04
Prostate cancer	10	3.40E-02	16	1.31E-05

Table 14: ***Down regulated pathways modulated in both species.***

	n° genes human	**FDR human**	**n° genes mouse**	**FDR mouse**
DOWN- regulated				
Focal adhesion	19	2.70E-09	22	5.98E-12
Pathways in cancer	17	4.61E-05	26	3.88E-11
Aldosterone-regulated sodium reabsorption	7	5.07E-05	6	2.53E-04
ECM-receptor interaction	9	5.41E-05	10	5.48E-06
Regulation of actin cytoskeleton	14	5.64E-05	12	3.18E-04
MAPK signaling pathway	15	5.96E-05	12	1.94E-03
Chemokine signaling pathway	12	1.51E-04	11	3.16E-04
Vascular smooth muscle contraction	9	2.92E-04	6	1.87E-02
Prostate cancer	8	3.03E-04	12	1.53E-07
Gap junction	8	3.10E-04	5	2.30E-02
Long-term depression	7	4.07E-04	5	1.11E-02
Glioma	6	1.94E-03	5	7.42E-03
Melanoma	6	2.97E-03	6	2.24E-03
Neurotrophin signaling pathway	7	1.05E-02	6	2.50E-02
Fc gamma R-mediated phagocytosis	6	1.17E-02	7	1.76E-03
Axon guidance	7	1.18E-02	11	3.88E-05
Endocytosis	8	1.71E-02	8	2.18E-02
Leukocyte transendothelial migration	6	2.00E-02	8	1.41E-03
Wnt signaling pathway	7	2.05E-02	6	4.82E-02
Colorectal cancer	5	2.42E-02	8	2.65E-04
GnRH signaling pathway	5	4.12E-02	6	8.76E-03

Since pathway analysis have indicated a excellent degree of similarity between the two cancer cell models as compared to their normal counterparts, next we decided to use, for a further analysis of modified pathways, only the common genes between mouse and human and identified by ANOVA as described in Table 11. As shown in Figure 27 the common genes between human and mouse were 499. To reinforce the analysis we decided to select from the 499 genes (~ 28% of 1774 human DEGs, see Table 11) only the genes showing a fold change ≥ 2 and ≤ -2. The resulting lists were respectively composed of 403 genes with a similar trend of expression and of 96 genes with an opposite trend of expression (Figure 27).

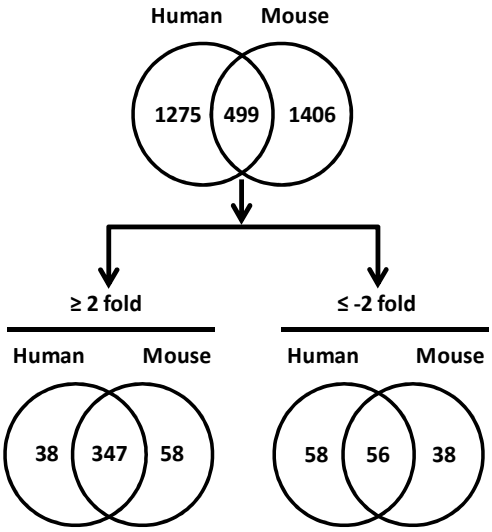

Figure 27: ***Venn diagram of the total genes identified by ANOVA analysis (Table 11)****. The identification of the genes with a similar and opposite trends, both in up-and down-regulated classes of human and mouse samples, are indicated.*

The Gene Ontology was performed on these two lists separately using Anduril tools (Ovaska et al., 2010). In particular, the visualization of enriched GO terms (Biological Process, Molecular Function and Cellular Component) of the 403 and the 93 genes, by using a p-value cut-off of 0.05 generated by Fisher's Exact Test, are shown in Figures 28 and 30 respectively. In both Figures each box is representative of a single GO term and its color ranges from white (statistically not significant) to red (highly statistically significant). Moreover inside each box has been shown the numbers of genes that belong to the category (f) and its p-value (p). Since not all the genes have a GO terms, the (f) number is lower than the initial number used as in puts (403 and 93). The thickness of the arrows indicates the principal direction followed by the genes in the tree diagram. In other words bigger thickness indicate a great number of genes following that arrows and hence that GO term.

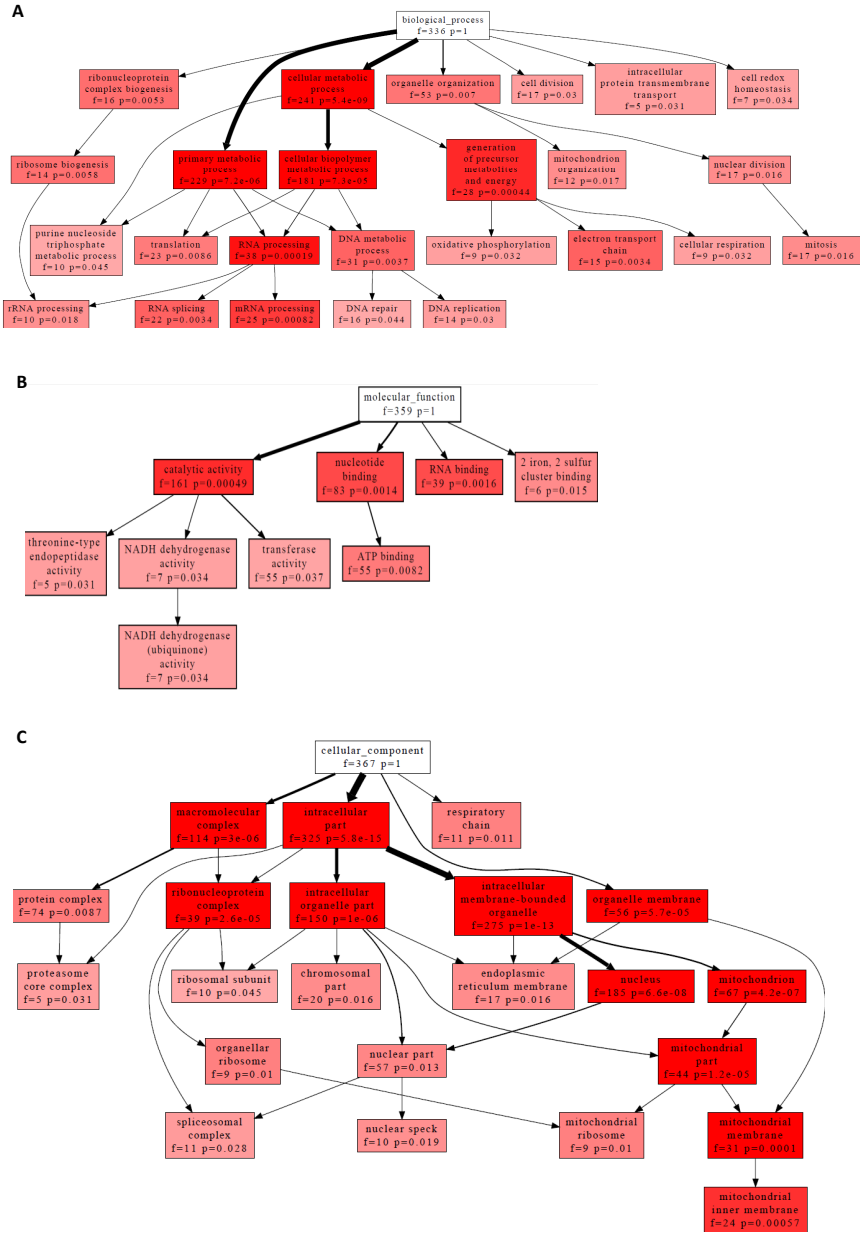

Figure 28: ***Gene Ontology analysis of 403 genes that change between the two species. A)*** *Biological Process,* ***B)*** *Molecular Function and* ***C)*** *Cellular Component. Enriched terms with p-value cut-off of 0.05.*

In Figure 28A is shown Biological Process. The 336 genes used as input identified two principal categories Cell metabolic process (~ 72% of 336

genes) and Primary metabolic process (~ 68% of 336 genes). Since from the top to the bottom of the tree diagram is possible observe an increase of the level of detail of GO term, we observed a significantly enrichment of the genes involved in RNA pre-processing (38 genes) and Generation of precursor metabolites and energy (28 genes).

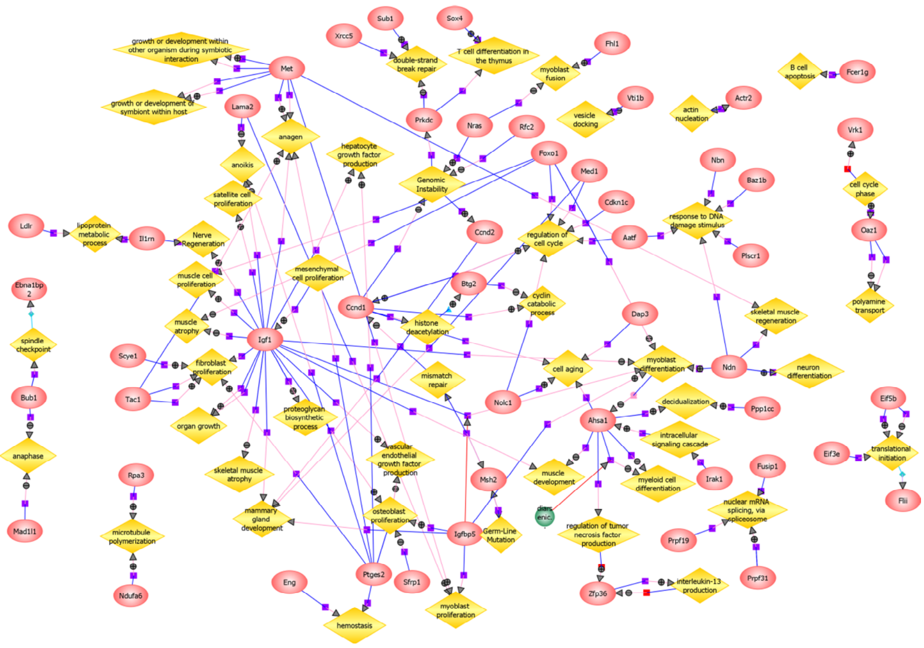

Figure 29: ***Protein-Protein Interaction enrichment analysis of the 403 genes that have a similar values of fold change in the two species.***

Another additional method used to obtain an informative graphic visualization of the process in which the 403 genes common between the two species is implemented in GeneSpring program. This program construct an interaction graph where nodes represent proteins encoded by such genes and edges correspond to the link with their specific GO Biological Process, as extracted from GeneSpring database able to integrate different sources (see Method). The Figure 29 shows Biological Process interaction investigation that is a layout of 52 proteins (~13% of 403 genes identified), 50 biological process and 141 PPI's. Integrative analysis of the results

obtained, described in Figure 28A and 29, indicated that the 403 genes are involved in specific biological processes like Cell cycle, Cell proliferation and RNA metabolic processes.

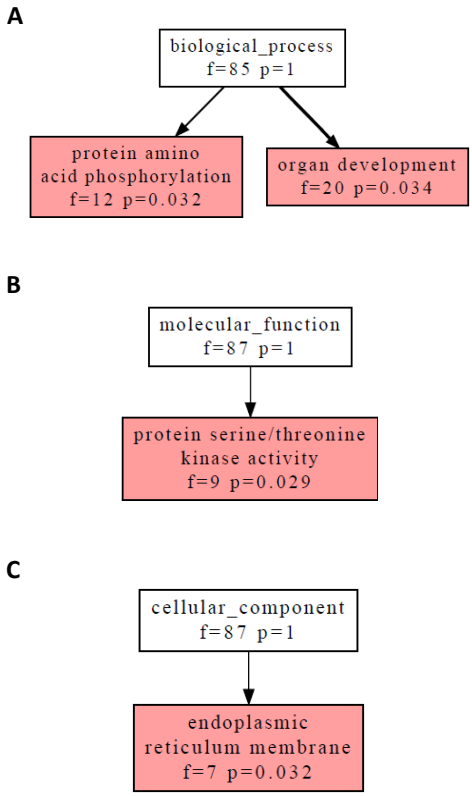

Figure 30: ***Gene Ontology analysis of 96 genes that change between the two species. A)*** *Biological Process,* ***B)*** *Molecular Function and* ***C)*** *Cellular Component. Enriched terms with p-value cut-off of 0.05.*

Since 96 genes is a small sample for this type of analysis, we did not expect to find a lot of information. However 12 out of 85 genes for which was possible to identify a biological process and 9 out 87 genes for which a molecular function was known, were related to a post-translational mechanism that is protein phosphorylation (Figure 30A-B). The majority differences between human and mouse are DEGs involved into environmental information, cellular communication and transport of

membranous system. Also KEGG pathways investigation was performed (data no show) that confirming the involvement of these genes in pathway as Gap junction, Focal adhesion and Leukocyte transendothelial migration (pathways).

Cross-species analysis from tissue to cancer cell lines, a nutshell of transformation

The 403 genes that correspond to common deregulated genes in both hCCL-hNT and in mCCL-mNT pair comparisons, thought to be essential for cancer development in all tumor cells analyzed, has been called *Nutshell*. In order to classify the Nutshell genes and identify group of genes able to discriminate between immortalization and transformation processes, the genes were divided in three categories by following criteria:

Shell: $DEG_i \sim$ mCCL vs mNT $\cap$ mICL-mNT $\neq$ mCCL-mICL

Middle: $DEG_i \sim$ mCCL vs mNT $\cap$ mCCL-mICL $\cap$ mICL-mNT

Nut: $DEG_i \sim$ mCCL vs mNT $\cap$ mCCL-mICL $\neq$ mICL-mNT

The *Shell* contains genes that were found important to passages from normal tissues to immortalized cell lines (the immortalization passage). The *Nut* contains genes that were found important to passages from immortalized cell lines to cancer cell lines (the transformation passage). And the *Middle* contains genes important both in the immortalization and transformation passages.

The results of CCL-ICL global comparison analysis indicated different distribution of fold change values on all chips. Therefore, in order not to lose significant information we have been use a distribution cut-off of 15% (of all probe set present on chip) to defined the significant values of fold changes for every classes (CCL, ICL and NT). Using this approaches we selected 253 genes for the Shell, 123 for the Middle and 27 for the Nut, see Figure 31.

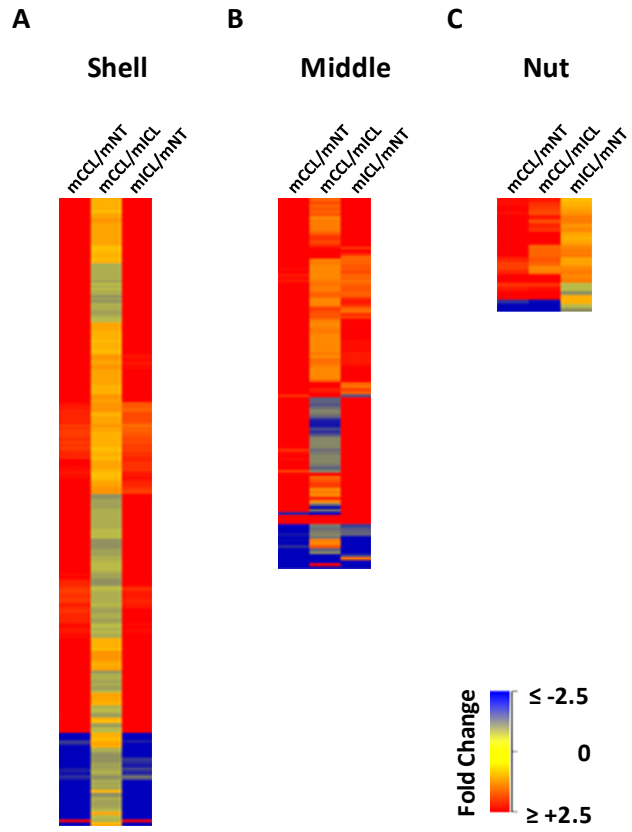

Figure 31: ***Hierarchical cluster of the 403 genes of Nutshell.*** *In figure are shown the* ***A)*** *deregulated genes of the Shell,* ***B)*** *the deregulated genes of the Middle and* ***C)*** *the deregulated genes of the Nut. The fold change values are color-coded as follows: red, up regulation; yellow, no change; blue, down regulation.*

Another level of investigation involves a Protein-Protein Interactions (PPI's) analyses that frequently are conserved through evolution. In order to extrapolate this information from the Nutshell, several data analysis programs were used, in particular the results have been obtained by using PINA (Wu et al., 2009) and GeneSpring (Rosenow et al., 2005) algorithms.

Both tools were performed to use different database for the extrapolation of protein-protein interaction as IntAct, BIND, BioGRID, MINT ect and different parameters can be selected to filter out redundant protein

interactions (see Methods). Starting from these different criteria to filter redundant information, the integration of the results obtained with the two approaches was performed.

The PINA analysis was performed by using 27 proteins as input list (Nut class). This analysis originated a layout of 590 proteins and 704 PPI's network. The GeneSpring analysis was performed by using the 403 genes of the Nutshell collection as in put list. This analysis originated a layout of 103 proteins and 142 PPI's. The results of both analyses then were used to build a layout network of 119 proteins and 131 PPI's networks (Figure 32A).

In Figure 32A each protein was represented as a colored circle as follow: green color -Shell-, brown color -Middle- and red color -Nut-. To identify the most important nodes in the network, a social network analysis was used, in particular, these nodes were selected using two principal criteria: the number of direct connections had to be > 25 and, if deleted, had to cause the isolation of a part of the network (a "broker" role in the network). In accord with both criteria 6 principal hubs were identified: HNRNPD, NONO, DHX9, CCND1, AHSA1 and PRPF4 (in Figure 32A are represented as colored boxes (Nut -red- and Middle -brown-).

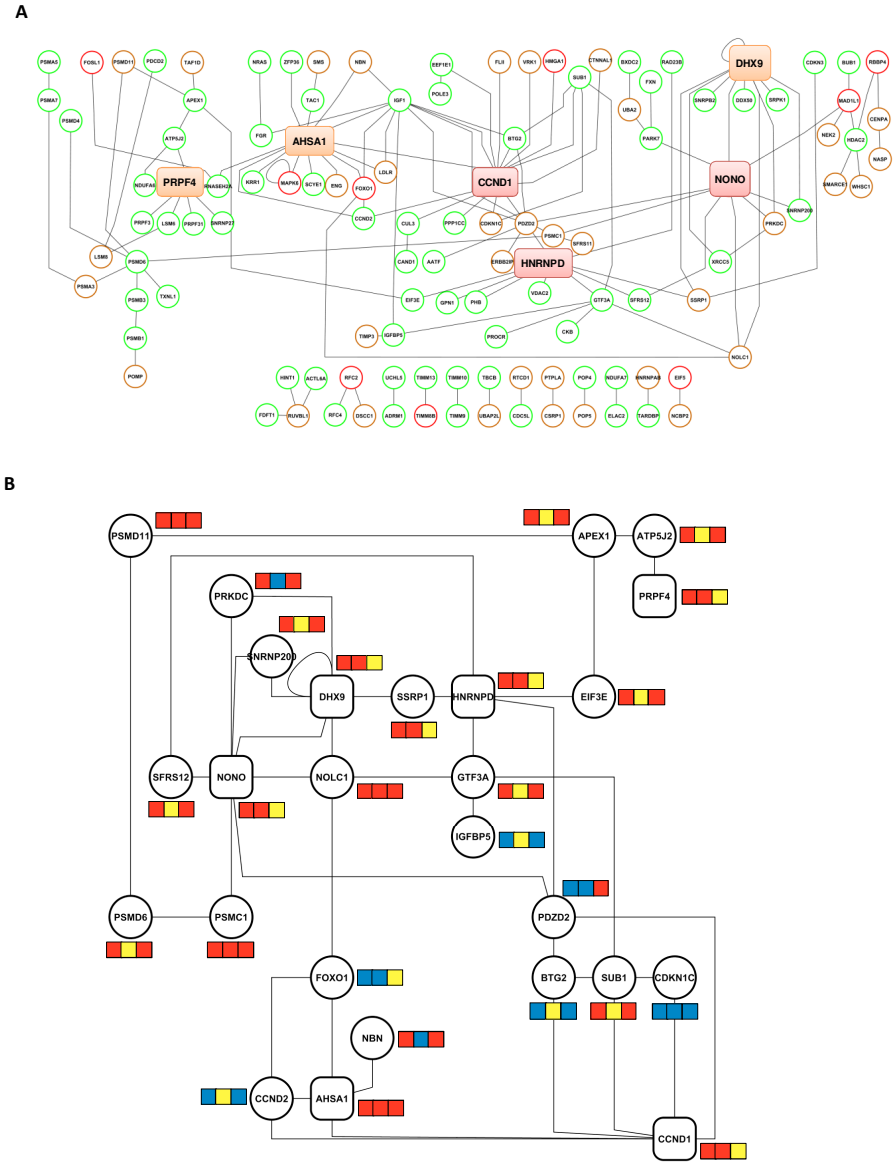

Figure 32: ***The predicted cancer circuit. A)*** *a layout network of 119 proteins and 131 PPI's networks. Green, brown and red circles are referred to Shell, Middle and Nut classes respectively and the colored boxes represent the hubs of the network. In* ***B)*** *has been shown a sub-network of the global network of (A) composed of 26 proteins and 40 PPI's. The colored boxes represent fold change of mCCL-NT, mCCL-ICL and mICL-NT comparison pairs respectively. The fold change values are color-coded as follows: red, up regulation; yellow, no change; blue, down regulation.*

Subsequently a sub-network were extrapolated using as parameter the more connected regions of the large interaction networks, it resulted on 26 proteins and 40 PPIs. The Figure 32B represents this sub-networks with the fold change values of CCL-NT, CCL-ICL and ICL-NT comparison pairs. These important result can be useful to shed light on a specific pathway of transformation that represent the minimum baseline matrix of transformation, independently of the organism under study (mouse or human).

6.1.2 Discussion

Global analysis of human and mouse datasets

The K-ras transformed mouse fibroblasts used as model of transformation, as indicated by this comparative analysis, well recapitulate the characteristics of different cancer human cells. In fact has shown by the comparative analysis performed by using several approaches (PCA, hierarchical clustering and Gene Ontology, Pathway and PPIs identification), many information recovered from the mouse model quite resemble that found in human models. In particular PCA analysis (accounting for 90% of variance in both species) show a whole transcriptome-scale similarities across species and samples. In fact, we could observe that the NT, ICL and CCL samples lay in different regions of the space and cluster separately, suggesting an high difference in their gene expression values. Moreover our data indicate that expression patterns across human and mouse are generally quite well conserved in classes NT, ICL and CCL, suggesting that the mouse cancer cell line may be a simpler and good model to study human tumors at least in terms of gene expression. Notable the mouse model of K-ras-dependent transformation has an enhanced similarity with human cancer cells harboring Ras mutations (i.e., CCRF-CEM, HL-60, MDA-MB231, ACHN, etc). This result was quite expected, but on the other hand, may indicate that the reduction of similarity between mouse and no K ras dependent human cancer

cells is the consequence of the different oncogenic mutations present in these human cancer cells. In fact, is well known that human cells to become transformed need to accumulate more mutations than mouse cells and that these mutations not always impinge Ras function. However the CCL-NT comparisons for both species have indicated that while both type of transformed cells are strikingly different from their normal counterparts, also indicated a strong similarity in terms of gene expression patterns between them. Notable, the vast majority of deregulated genes identified in the transformed cell lines were up regulated, suggesting that the transformation process induces a general increase of gene expression. This observation was confirmed also in the comparisons between ICL-NT, revealing that also immortalization induces a general increase of gene expression (data no show).

The ability of mouse model to resemble human models has been strongly supported also by pathways analysis. In fact about 72% of the most statistically significant pathways overlapped between the two species.

Cross-species analysis from tissue to cancer cell lines, a nutshell of transformation

Our comparative analysis permitted also the identification of 403 common genes altered in all hCCL-hNT and mCCL-mNT comparison pairs. Such genes, namely Nutshell, represent the most significantly altered genes across all human tumor cells and across the two species. We have classified these genes in 3 categories, in order to model the flow of gene expression from normal tissues to immortalized cell lines and finally to cancer cell lines. The underlined idea is that each step driving normal cells to immortalization and then to transformation can be characterized by some specific gene pattern. In this regard the three following categories has been recognized: Shell, genes important for the transition from a normal tissue to an immortalized cell line (253 genes); Nut, genes important for the transition from an immortalized

cell line to a cancer cell lines (27 genes); Middle, genes important for both transitions (123 genes). This classification direct us to consider the nutshell as the baseline matrix of transformation, independently on the system under study. Moreover using different approaches for information extrapolation as protein-protein interaction and GO enrichment etc, the information enclosed into the Nutshell was analyzed permitting the definition of a cancer circuit formed of 119 entities (proteins) and 131 interactions. Afterwards, using a social network analysis, we identified in this circuit 6 main hubs that revealed a sub-network (called *minimum cancer circuit),* composed of 26 entities and 40 interactions. These hubs or crucial nodes are HNRNPD, NONO, DHX9, CCND1, AHSA1 and PRPF4.

Interestingly, some of these genes have already been linked to tumorigenesis i.e. CCND1, but not always well studied i.e. HNRNPD (Fawal et al., 2006; Gouble et al., 2002), NONO (Salton et al., 2010) and DHX9 (Zucchini et al., 2008). Thence these genes can be considered a fruitful earth to future investigations in order to understand the genetic common baseline of cancers.

6.2 Promoter Scan: Algorithm to detect over-represented TFBSs in the proximal promoter regions of co-regulated or co-classified.

Among the most fascinating open questions in biology today are those associated with the global regulation of gene expression associated to the execution of the vast majority of cellular processes. The answers to some of these questions have been moved from few steps closer to realization with the advent to DNA hybridization microarrays. DNA microarrays generate large amounts of numerical data that give the possibility to monitor the mRNA expression levels of thousands of genes at same time point (Lockhart and Winzeler, 2000). For instance, a primary goal in the analysis of such large data sets is to find genes that have similar behavior under the same experimental conditions. Several clustering algorithms are available to group

genes that have a similar expression profile (Altman and Raychaudhuri, 2001; De Smet et al., 2002; Eisen et al., 1998; Heyer et al., 1999; Tavazoie et al., 1999). Given a cluster of genes with highly similar expression profiles, we can search for the mechanism that is responsible for the coordinated behavior of the genes that belong to the cluster. We basically assume that co-expression frequently arises from transcriptional co-regulation. As co-regulated genes are known to share some similarities in their regulatory mechanism, possibly at transcriptional level, their promoter regions might contain some common motifs that are binding sites for transcriptional regulators (Brazma et al., 1998; Wolfsberg et al., 1999). The main difficult of this investigation is the identification of the TFBSs (or motifs) into a promoter region, that can be compared to the search of a needle in a haystack because these sites are short nucleotide sequences (typically 6-20 nucleotides) inserted in the midst of a great amount of statistical noise (a typical input being one regulatory region of length 1000 bp upstream of each gene). To make matters worse, there is sequence variability among the binding sites of a given transcription factor, and the nature of the variability itself is not well understood. Over the past few years, several tools have become available for motif prediction as MotifSampler (http://bayesweb.wadsworth.org/gibbs/gibbs.html), AlignACE (http://arep.med.harvard.edu/mrnadata/mrnasoft.html), MEME (http://meme.sdsc.edu/) ect. A guidance to users regarding the accuracy of currently available tools in various settings is reviewed in (Tompa et al., 2005).

As previously mentioned, the basic problem of several promoter prediction algorithms is the high number of TFs and TFBSs identified. However, while the identification of a large number of TFBSs for a given TF into a promoter region permit a better estimation of its binding probabilities, on the other hand a high number may increase the noise into the promoter

prediction model. We approached the problem addressing mainly two aims. The first aim of our algorithm has been the reduction of the number of TFs predicted into a promoter region by identification of more biological relevant TFs. The second aim has been the set up of a procedure able to build a promoter model from a given set of genes. Our Promoter Scan algorithm, validated using a set of co-regulated genes extracted from literature data, shows that the number of predicted TFs may be decreased in accord with the increase of new combinatorial parameter namely Signal Transmission Intensity (STI), that permits a strong reduction of predicted TFs (noise) without reduction of the "true" TFs.

6.2. 1 Results

Promoter Scan Algorithm

The algorithm is based on two well known discriminative concepts of transcription factors identification studies, that are the presence of multiple copies of a specific TF in a promoter region, often observed in higher organisms, and its phylogenetic conservation. Starting from these widely accepted assumptions, a goal of our work has been the development of an algorithm able to extrapolate from the complex landscape of predicted TFs an essential set of them necessary to the function of a composite proximal gene promoter region. Moreover, since exists a strong similarity between genome data of closely related organisms, we used a phylogenetic footprinting model, rather than relying on the vast intergenic wastelands, to significantly increase the sensitivity of our method of identification of putative TFs. In fact, given that has been shown that nucleotide variations within functional regions of a gene accumulate more slowly than variations in regions without sequence-specific function (Blanchette and Tompa, 2002; Monsieurs et al., 2006; Sumiyama et al., 2001; Van Hellemont et al., 2005; Woolfe et al., 2005), the comparison between sequences of homologous genes could help to identify their transcriptional regulatory DNA portions.

The underlining assumption is that the transcriptional control of homologous genes is under the control of similar and conserved mechanisms (McCue et al., 2002; McGuire et al., 2000; Wasserman et al., 2000). Given these statements, the algorithm has been implemented by hypothesizing that the regulation of homologous genes has been subjected to a stronger evolutionary pressure than the surrounding sequences, preserving their specific DNA sequence but not their relative position inside the promoter region. In addition, to avoid the difficult to have a good prediction model of long promoter sequences, the algorithm has been developed to identify TFs in a very close region surrounding the Transcriptional Starting Site (TSS), called proximal promoter region, spanning between -500bp to +199 bp. Finally, we have evaluated the ability of our algorithm to construct a statistical promoter structure model by using a set of co-regulated genes as compared to a great number of randomly selected genes. This comparative analysis strongly supported the validity of our approach.

Algorithm implementation

A) The Signal Transmission Intensity (STI)

To start, $\forall j$ where j identifies a human proximal promoter region, we have a vector of elements i (that represent each putative TF) and for each predicted TF_i the probability $P_i(j)$ is estimated on the basis of the relative frequency of occurrence of each TF in the promoter region considered. Then we define for each TF_i on the promoter region

$$P_i = \frac{m_i}{M} \qquad \textbf{(Eqn.27)}$$

where m_i is the number of occurrences of ith TF and M is the total number of TFBSs for each TF. Furthermore, we can define the Entropy of TF_i as the probability of getting itself under a jth promoter region. That is

$$H_i = P_i \log_2(P_i) \qquad \textbf{(Eqn.28)}$$

We assume that the probability to binding of *i*th TFs is influenced by the presence of other TFs on the promoter region. In other words, we assume that the signal of specific motif for each TFs present in the proximal promoter region is determined also by the contribution of all TFBSs present on the promoter. In this way the entropy H_i gives a measure of the chaotic structure of the promoter (uncertainty or information).

Moreover, we introduce a variable *w* which is the contribution of *i*th TFs to the structure complexity of a promoter of length *L*. In this way, we introduce the concept of Amplitude Transmission Intensity (ATI) by TFBS to its specific TF as

$$w_i = \log_2\left(\frac{(m_i\, l_i)}{L}\right) \qquad \textbf{(Eqn.29)}$$

where *L* is the length of promoter region in bp unit, *m* is the number of occurrences and *l* is the length of specific TFBS for each TF_i in bp unit. In this way, we can assume that ATI(*i*) is a monotonic decreasing function of *L* if we consider *ml*(*i*) as constant

$$\mathrm{Lim}_{L\to\infty}\, w_i = 0 \qquad \textbf{(Eqn.30)}$$

This means that the increase of the length of the promoter region induces a dispersion of the information and consequently a reduction of the information obtained by increasing the amount of noise within the promoter region where the TF_i can be hidden.

The variable ATI is used together with the variable H_i in order to define the intensity and the amplitude of the information of *i*th TF on promoter region *j*, with

$$I_i = -w_i H_i \qquad \textbf{(Eqn.31)}$$

where *I* is called the Signal Transmission Intensity (STI). The new parameter STI, giving in the same time frequency (H) and amplitude (ATI) of

information, permits to evaluate the transmitter information of the TF_i from the promoter region to the outside to guide TFs towards the promoter.

B) Phylogenetic Source Conservation (PSC)

In some circumstances, the only alignments between promoter regions may not be so informative, in fact the alignments may bear little relationship as compared to the underlying conserved configurations of TFs inside the promoters (Lenhard et al., 2003; Santini et al., 2003). Starting from this observation, we supposed that in the comparative procedure between homologous genes pairs, the alignment of sequences was less informative than the global conservation of the promoter structure. Therefore, we introduce a Phylogenetic Source Conservation (PSC) parameter that estimates how the information regarding a TF_i obtained in a principal species (in our case human) is conserved in a reference species (in our case mouse). To this end, we recovered for every *j*th human promoter region a *k*th homologous promoter region from mouse and we computed the STI for the same *i*th TF in this species. The PSC is then calculated using the following formula:

$$PSC_i = 1-|I_h - I_m| \quad \textbf{(Eqn.32)}$$

where I_h and I_m are the Signal Transmission Intensity obtained by *i*th TFs on proximal promoter region of human and mouse, respectively and PSC represents the Phylogenetic Source Conservation.

C) The STI and PSC combination and the minimum transmitter information filtering

The combination of the Signal Transmission Intensity (STI) and the Phylogenetic Score Conservation (PSC) gives the Evolutionary Transmitter Information (ETI) for every *i*th TF on structure of *j*th promoter region, and it is introduced by new variable β, as

$$\beta_i = PSC_i * I_i \quad \textbf{(Eqn.33)}$$

Then $\beta(i)$ is a monotonic function of PSC and

$$\mathrm{Lim}_{PSC \to 1}\ \beta_i = I_i \qquad \textbf{(Eqn.34)}$$

The resulting vectors for each *j*th promoter region will be like

$$V_j = \{\beta_1, \beta_2, \ldots, z\}$$

where z correspond to each TFs. In order to filter out the TFs with ETI close to a background of promoter sequences, we applied a cut-off to each element of V_j, assuming that the TFBSs showing a low ETI value do not drive their respective TFs to the promoter region.

D) Promoter model identification

A set of genes can be grouped into clusters according to their similar expression profile or their similar function (Hughes et al., 2000; Tavazoie et al., 1999). The identification of co-occurrence of TFs in a group of gene promoter regions is one of the fundamental steps to describe biological phenomena in a more accurate way. Several TF detection algorithms have been developed to discover overrepresented motifs in a set of co-expressed genes (for instance (Bailey and Elkan, 1995; Hertz and Stormo, 1999; Lawrence et al., 1993; Liu et al., 2001; Thijs et al., 2002; van Helden et al., 1998; Workman and Stormo, 2000)). The rationale behind these methodologies is that a set of genes regulated by the same TF should contain a statistically overrepresentation of the binding motif for such a TF as compared to its occurrence in unrelated genes. By using our algorithm, we start from a list of co-regulated or co-classified genes (D) and for each proximal promoter region of this gene list we computed a global matrix

$$\exists\ K = \begin{pmatrix} \beta_1 & 0 & 0 \\ \beta_1 & \beta_2 & \beta_3 \\ 0 & 0 & \beta_3 \end{pmatrix}$$

where rows represent a vector V for the *j*th promoter region. Starting with this matrix we build a new vector that represents the promoter model

containing the total identified TFs present in all the promoter region of the set D, but with the constraint that each TF have to be present in at least 75% (co-occurrence percentage) of the promoter regions listed in D. The resulting TFs set is believed to participate in regulating gene transcription of the gene list D. The resulting representative promoter vector is then building as

$$V(D) = (0 \ \beta_2 \ 0 \ \beta_4 \ \beta_5 \ \beta_6 \ \ldots)$$

where β represent a new STI calculated on all the promoter regions present in the list D.

E) Permutation background control

Since one or more predicted TFs in a promoter region could be only background noise and that the new STI value calculated for the *i*th TF could be close to the value of STI calculated for the entire genome (background), we performed a statistical test to estimate the significance of *i*th TF in the promoter region model *V(D)*, by calculating the STI value for the same TF in a set of random promoter region models. In other words, we tested the hypothesis that the *v*th element of *V(D)* has a higher information (that is a higher STI value) in the *V(D)* model as compared to a random *V'(D')* model. We tested a series of rearrangements by taking a set of random promoter regions from the human genome (*V'*, Background Set), with the same dimensional number of *V*. We have calculated the probability that we could observe an equal or greater β value if the predicted TF is selected randomly, given by

$$V(D) = \begin{cases} 1 \ where \ V(D) \leq V'(D') \\ 0 \ where \ V(D) > V'(D') \end{cases}$$

The p-value is calculated for *V(D)* = 1 by counting the number of cases in which the alternative hypothesis is true

$$P(V) = \frac{\sum V(D)}{N} \qquad \textbf{(Eqn.35)}$$

where *N* is the number of comparisons observed and where a small p-value corresponds to a great statistical significance. We evaluated the biological relevance of each predicted TF using the p-value and constructing a new statistical promoter structure model with our approach.

Promoter Scan validation by using a set of known gene promoter structure

In the following paragraphs, we address the problem of the potential false positive TFBSs filtering by using ten different group of genes as test (Test Set), in which at least one TFBS for each group has been experimentally well characterized (i.e., by Electrophoresis Mobility Shift Assay, functional analysis and Northern blot). Then we give an overview of the output returned by the algorithm and the meaning of the resulting scores. Finally, we discuss the identification and validation of a promoter structure model for each group, in order to show the ability of Promoter Scan to build a statistical promoter structural model of a gene cluster in which the p-values generated by the algorithm as outcome are derived from the comparative analysis between the Test Set and a random set (Background Set).

Table 14 shows the Test Set on which the algorithm has been used. Specifically, the test set comprises 10 groups of genes (namely ID and indicated with a roman number) with a variable dimension (N, number of genes considered for each group). In addition in Table 14 has been also indicated the TF common to all the genes of each group (Genes) and experimentally validated (TF known). The test set, selected in this manner, contained 228 human promoter gene regions. In order to estimate the sensibility of our algorithm, each group of the test set has been used to evaluate the systematic effect of potential TFs false positive (FP) reduction on the amount of true positive (TP) TFs. Afterward, to evaluate the significance of the predicted TFs in the Test Set, we applied the algorithm to a set of 1000 gene promoter regions taken randomly from the human genome (Background Set) as background (data not shown). Since the

algorithm works using also a phylogenetic score, for every human promoter region present in the Test Set and the Background Set, we identified the corresponding mouse homologous gene promoter region. Since many regulatory regions controlling transcription rate are proximal to the TSS, we retrieved in both homologous human and mouse promoter regions only the nucleotide sequences close to TSS. Specifically we analyzed the region between 500 bp upstream of the TSS and 199 bp downstream from it. This 700 bp region, that can be a good estimation of the real proximal promoter region, being relatively small, permitted the prediction of a relative low number of TFBSs.

As transcription binding sites collection, we took advantage of the JASPAR database (http://jaspar.genereg.net/) from which we recovered 79 human TFs and their relative binding sites (TFBSs). Subsequently, the human and homologous mouse promoter regions (corresponding to the 1228 promoters present in Test Set and Background Set) were submitted to MotifMatch algorithm implemented on Anduril framework (Ovaska et al., 2010) and the resulting matrix of occurrences (97012 elements) was used as input to further analysis with Promoter Scan algorithm.

A) Selection of highly informative TFs

MotifMatch algorithm analysis of the human promoter regions of Test Set, applied with default parameters, identified an average of ~ 51 predicted TFs for each promoter region on the total of 79 TFs used as input (35.5% of reduction). As previous described, we first evaluated the ability of our algorithm to reduce the potential false positives TFs. In particular, for each promoter region of the Test Set and Background Set, we calculated an ETI value for every TF present in the two Sets. This operation gave as results more than 60000 ETI values. From this list of ETI values have been selected and graphically represented only the ETI values corresponding to the 10 TFs (true TFs) shown in Table 15 (Test Set) as compared to Background Set.

Table 15: ***The Test Set database****. For each group, are represented the name of Group (ID), the TF experimental validated (TF known), the number of promoter selected (N) and the name of the genes to which the promoter belong (Genes).*

ID	TF known	N	Genes
I	AP1	20	TNFAIP6 TIMP1 TGFB1 TFRC TFF1 SPRR1B SPP1 SERPINB2 PENK OXTR NTS MTCH1 MMP9 MMP7 IVL GJA1 ETS1 CD80 CCL5 CCL2
II	CREB1	30	MYST2 SPATA2 STAT3 FN1 MOG TIMP1 DBH FOS IGFBP1 NPC1 POLD2 ADRB2 POLB TGFB1 BRCA1 CFTR ERBB2 EXO1 GNRH1 IFNG NR3C1 HPGD NF1 ODC1 PPARA SLC5A5 ACHE CCND1 JUN PCDHB11
III	ESR1	19	APOE AVP CAD CCND1 CDKN1A CXCL12 DFFB EGFR FOS GADD45A HOXA10 HSPB1 IFNAR1 JUNB KRT19 LTF PELP1 PPARG STAT5A
IV	ETS1	30	ANGPT2 ANPEP BMP4 BRCA2 CD4 CD79B CD8A CSF2 CSNK2A1 DNTT EGR1 ETV4 GRPR HMOX1 IL12B IL3 IL5 LTB MMP3 MPL NOS3 PRL SLC26A3 THBD TIMP1 TNC TNF TNNC1 VWF WAS
V	NFIC	30	ADA ALAS1 ARG1 COL18A1 CSNK2A2 CYP11A1 CYP1A2 CYP27B1 ELN GFAP GNRH1 GTF3C1 GUCY1B3 HMOX1 HSD11B2 HSPA1A MBP NPY PCK2 PFKFB1 PKLR PLAT POLD2 PRG2 SFTPB SFTPC SLC34A2 SMUG1 SPATA2 VWF

VI	NFKB	15	PTX3 PDGFB MMP9 MAPK14 MAP3K8 LTB ICAM1 HSD11B2 ELF3 CXCL1 CSNK2A2 CSF3 CSF1 CD83 B2M
VII	SP1	30	ACACB ALDOA ALOX5 CAT CCND1 CDKN2D COL18A1 COL7A1 CXCL1 EGFR ESR1 EXO1 FOSL1 GLP1R HNF4A HOOK2 HSD17B1 HSPB1 KIT KRT16 MAT2A MMP14 MPG PAFAH1B1 PDHA1 POLB SLC5A1 STAR TGFB1 TNFRSF10B
VIII	STAT1	10	VIP PTGFR PRF1 IL6ST IL2RA IFNG FOS CCL2 BCL6 A2M
IX	TFAP2A	35	ACHE ADM ALPPL2 CALB2 CDKN1A CFTR CHGA CYP11A1 DBH ERBB2 FBLN1 GFAP HK2 HMOX1 HOOK2 HSD17B1 HSPA1A MAPT MCAM ME2 MPG MYC MYO6 PIM1 PLAT POLD2 PTGDS RGL3 SLC19A1 SYT1 TALDO1 TH TIMP1 TIMP2 TNPO1
X	USF1	9	APEX1 CALCA CEACAM5 CEL GCK HMOX1 IGF2R KCNN3 MYH9

Figure 33 shows the curves obtained by using the ETI values of these 10 TFs for both Test Set (continuous line) and Background Set (dashed line). In particular, on the x-axis the TFs are ranked relatively to their ETI content increase, and on the y-axis are represented as ETI value. As shown in the Figure the true TFs of the Test Set have a higher discriminative curve profile as compared to Background Set curve profile. Indeed the Test Set curve permit to distinguish the true TFs from the noise data, represented by the curve of Background test, because already at 5% (see the box) of the Test Set curve the ETI values are higher than the Background Set (ratio ~ 15),

corresponding to a discriminatory value between 0.001-0.002. Such a discriminatory value was used as cut-off to discriminate between real TFs and noise. Cut-off application to our data give as result an average of ~ 33 predicted TFs for each promoter region analyzed. In other terms our algorithm brings to a 57.7% reduction of the putative TFs (used as input), without loss of true positive TFs.

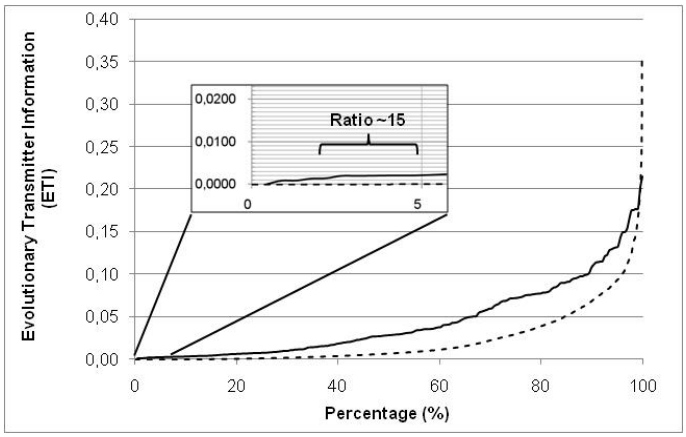

Figure 33: ***Evolutionary Transmitter Information (ETI) distribution of true TFs in the Test Set (-) and in the Background Set (--).*** *In the small box is shown the region of the graph between 0% and 5% in which the ETI value of Test Set is already higher than the Background Set (Ratio ~ 15).*

Taken together these results demonstrate that the our estimated ETI value is able to discriminate from the true TFs and noise data, in accord with the biological postulated evolutionary multi-copy effect of real TFs. Therefore the ETI is an accurate parameter value to eliminate false positive present also in the Background Set.

A) Identification of statistical promoter structure model

In this section, we evaluate the ability and stability of Promoter Scan algorithm to build a statistical promoter structural model for each group under study. Our algorithm has been implemented in a way that the user may decides the minimum ETI, the minimum number of sequences in which the searched TF has to be present and the number of permutation tests to

perform. We estimated that Promoter Scan works properly by using a cut-off ≥ 0.001 for minimum ETI and the predicted TF have to be present in at least 75% of the input gene list (minimum number of sequences).

For each group, the output of the algorithm is a common pattern of TFs that is the structure of the promoter region model, filtered out by noise TFs. Such a promoter structure model, comprising only TF shared in at least 75% of initial input, is used to recalculate for each TFBS a new STI. The significance of these new scores is evaluated by 10000 permutation tests that will bring to a final output in which for each selected TF the algorithm reports a minimum, a maximum, the median and the mean of p-values collected after the permutation tests (Table 16, results obtained using the Group I of Table 15). In this example it is clearly shown as Promoter Scan is able to perform a strong reduction of initial input (~ 60% of the 79 human TFs present in our collection), value that can be further reduced by applying cut-off on the p-values.

Table 16: ***The list of p-values obtained by application of Promoter Scan algorithm on Group I.*** *The column "Potential TFs" represents the predicted TFs present in the promoter structure model. The other columns report the min, the max, the median and mean p-values obtained from the 10,000 comparisons with the Background Set rearrangements. The experimental validated TF, AP1, present in all the promoter sequences of Group I is shown in italic.*

Potential TFs	**min**	**max**	**median**	**mean**
AP1	0.000	0.000	0.000	0.000
FOXC1	0.000	0.060	0.020	0.021
GATA3	0.000	0.030	0.000	0.004
NFATC2	0.000	0.020	0.000	0.001
SOX10	0.000	0.020	0.000	0.002
YY1	0.000	0.040	0.005	0.007
FOXD1	0.010	0.110	0.060	0.058
STAT1	0.010	0.120	0.060	0.061
GATA2	0.020	0.140	0.060	0.062
HLF	0.020	0.180	0.100	0.103

NR4A2	0.030	0.160	0.090	0.090
SOX9	0.030	0.190	0.090	0.098
FOXL1	0.040	0.220	0.120	0.123
HOXA5	0.040	0.190	0.110	0.117
Pax6	0.040	0.260	0.120	0.117
SPIB	0.110	0.350	0.220	0.218
REL	0.120	0.310	0.200	0.206
ESR2	0.150	0.340	0.250	0.251
SPI1	0.170	0.410	0.275	0.274
SRY	0.170	0.350	0.230	0.241
ETS1	0.190	0.480	0.295	0.291
ELK1	0.280	0.490	0.385	0.385
MZF1_1.4	0.330	0.540	0.425	0.426
FOXA1	0.480	0.720	0.600	0.599
ZNF354C	0.560	0.780	0.690	0.684
BRCA1	0.600	0.800	0.670	0.674
CREB1	0.620	0.810	0.730	0.722
ESR1	0.760	0.940	0.850	0.851
SP1	0.760	0.940	0.870	0.871
MZF1_5.13	0.910	1.000	0.970	0.966
NFIC	0.940	1.000	0.980	0.978
TFAP2A	0.990	1.000	1.000	1.000

In Table 17 we report the p-values obtained by applying Promoter Scan for the true TFs for each group present in the Test Set. As observed they range between 0-0.06. Based on these p-value range we decided that the predicted TFs can be considered as biological relevant if at least three of the four collected p-values (minimum, a maximum, the median and the mean) are $\leq$ 0.05. Applying this criterion stringency, the most relevant TFs selected for every proximal promoter region are ranked between 2-8 (see Table 18) that means a reduction of about 90-97% of the initial TF human collection used as input. The application of these p-value cut-off caused the lost of STAT1 transcription factor in the group VIII, indicating that less of 5% of real TFs has been filter out upon application of strong constraints.

Table 17: ***The list of p-value obtained by application of Promoter Scan algorithm on the complete list of Test Set.*** *For each group (ID) are reported the TF experimental validated (TF known) and the resulting p-values for each of them.*

ID	TF known	min	max	median	mean
I	AP1	0.000	0.000	0.000	0.000
II	CREB1	0.000	0.030	0.000	0.004
III	ESR1	0.000	0.020	0.000	0.001
IV	ETS1	0.000	0.050	0.010	0.013
V	NFIC	0.000	0.000	0.000	0.000
VI	NFKB	0.000	0.000	0.000	0.000
VII	SP1	0.000	0.000	0.000	0.000
VIII	STAT1	0.000	0.060	0.010	0.014
IX	TFAP2A	0.000	0.010	0.000	0.000
X	USF1	0.000	0.000	0.000	0.000

In order to validate our approach we compared our computational identified TFs with that experimentally validated and published. In particular, as shown in Table 18, for each group of promoters described in Table 15 we identified several TFs. Such a comparative analysis indicated that some of our TFs were already identified in the same promoters (Table 18, indicated in bold), again confirming the strength of our approach.

To further assess the specificity of our algorithm, we performed the complete procedure on a collection of 10 randomly groups of gene promoters (Random Set). In each of them no significant TFs have been identified, confirming the robustness of Promoter Scan to filter out the false positive data (noise). The results of application of Promoter Scan on Test Set indicate that the STI can contribute to the characterization of the promoter region of co-regulated genes as well co-classified genes. This contribution is mostly obtained through the substantial reduction of the overwhelming number of candidate TFs and the built of statistical promoter structure model.

Table 18: ***The list of TFs identified in the Test Sets.*** *Each row represents the promoter structure model resulting by Promoter Scan algorithm (p-value cut-off 0.05). The transcriptional factors identified in literature (experimentally validated) and identified by Promoter Scan are shown in bold.*

ID	TF known validated	Potential new TFs
I	AP1	GATA3; NFATC2; SOX10; YY1
II	CREB1	**ESR1**; MZF1_1.4
III	ESR1	MZF1_1.4; **SP1**; ZNF354C
IV	ETS1	NFATC2; YY1
V	NFIC	**ESR1**; ESR2; MZF1_1.4; **SP1**; ZNF354C
VI	NFKB	MZF1_1.4; **SP1**
VII	SP1	MZF1_1.4; MZF1_5.13; **NFIC**; **TFAP2A**; ZNF354C
VIII	-	**REL**; YY1
IX	TFAP2A	**ESR1**; MZF1_1.4; MZF1_5.13; **NFIC**; NHLH1; **SP1**; ZNF354C
X	USF1	ESR1; ESR2; **NFIC**; NR4A2

6.2.2 Discussion

The proposed algorithm enable to identify the most biological relevant TFs into a clusters of co-regulated or co-expressed gene promoters by using two new parameters, Signal Transmission Intensity (STI) and Phylogenetic Source Conservation (PSC). The work presented here provides a computational framework for the identification and the modeling of regulatory sequences present into a promoter on the bases of their conservation evolution. The two key aspects of this algorithm are: first, the ability to extrapolate the intrinsic information of each TF present into a promoter region and second, the modeling of the significance structure of proximal promoter.

The complete workflow is shown in Figure 34: starting from a single proximal promoter region for each considered gene of the cluster, and a list of all its putative TFBSs, we may evaluate the transmitter information of

each TF computing the Signal Transmission Intensity (STI) that is given by the frequency (H) and the amplitude (ATI) of the information. The STI represents the intensity and the amplitude of the signal for each TF in a specific proximal promoter region, that in other word we can consider as the transmitter information of the TFs from the promoter to the outside (Step 1). Then for each human gene considered, we have to identify the mouse homologous. The STI value computed for the promoter region of the mouse homologous is compared to STI of human promoter in order to give to the human promoter a Phylogenetic Source Conservation (PSC) score (Step 2). Subsequently, the combination of the Signal Transmission Intensity (STI) value and the Phylogenetic Score Conservation (PSC) value, produces a new parameter namely Evolutionary Transmitter Information (ETI) value. This parameter introduces the first level of noise filtering for predicted TFs because it allow to the exclusion of TFs that are not perceived outside of promoter region. In the end, given a list of proximal promoter regions passed through the previous steps, we can select the hit TFs in the group of co-classified or co-regulated genes, applying a permutation test against a random Background Set (step 3). Altogether these different steps permit to build a statistical promoter structure model that well describes the putative common promoter of the genes cluster considered in the analysis.

In this work, we showed that our Promoter Scan algorithm is able to find overrepresented TFs in several well-described Test Sets of proximal promoter sequences. We focused on the influence of different parameters on the performance of the algorithm (i.e., ETI). The Promoter Scan algorithm has been validated by using 10 sets of promoter sequences in which one or more regulatory elements have been already experimentally validated by other authors. These Test Sets allowed us to quantify, up to a certain level of confidence, the performance of our algorithm. From a biological point of view, it is very interesting to note that the algorithm is able to find

statistically significant combinations of TFs in a group of co-expressed genes, reducing the number of candidate TFs and hence leading to a statistical promoter structure model. The results of this study may lead to a new strategy of promoter analysis that can be more useful for understanding of regulation of gene expression as compared to other algorithms having the same function.

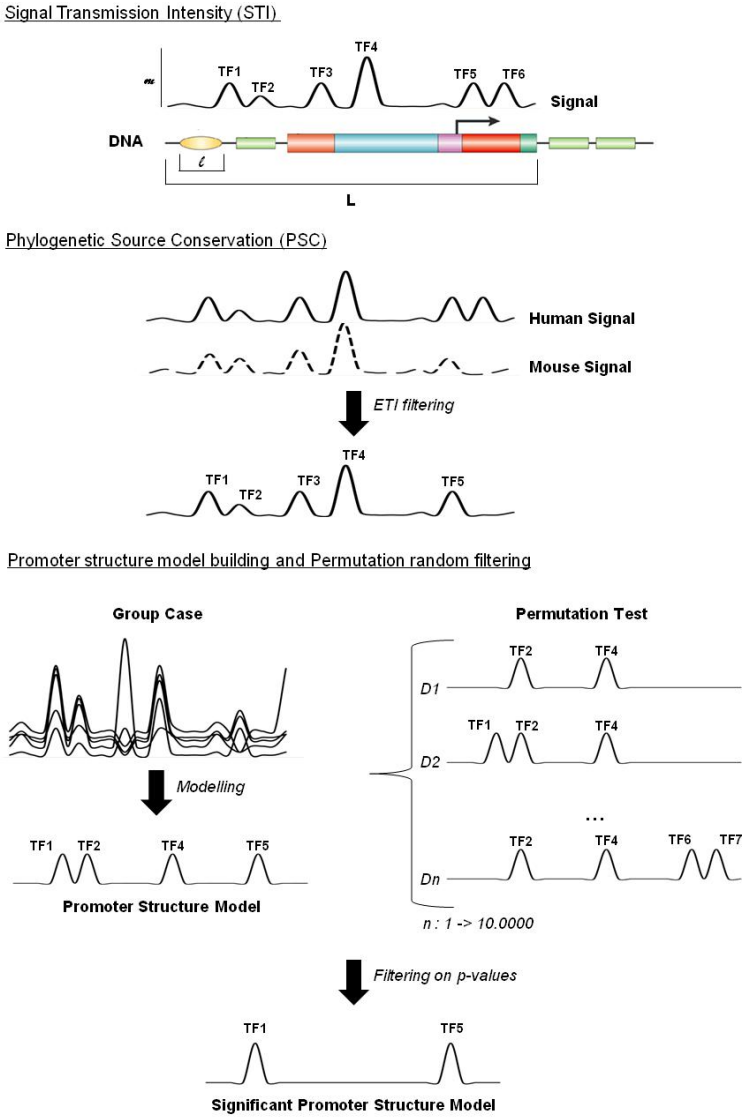

Figure 34: ***Flowchart of Promoter Scan algorithm.***

Materials and Methods

7.1 Gene expression profiles comparative analysis of immortalized and K-Ras transformed mouse fibroblasts grown in different glucose availability.

Experimental design

In order to identify the transcription profiles of normal and transformed mouse fibroblasts grown with different initial glucose availabilities, we decided to modulate carbon metabolism by using two different initial glucose concentrations: high (25 mM) and low (1 mM). Cells were followed for at least 96 h, i.e. from the moment of seeding to when they either reached confluence or started to grow in multi strata or to die. The cells, to be used for the transcription analysis, were collected at 16 h from the initial seeding, time corresponding to the change of medium (25 and 1 mM glucose, indicated as T0). Then both cell lines, grown in 25 and 1 mM glucose, were collected at 24, 48 and 72 h and prepared for the transcriptional analysis as reported (see experimental design in Figure 12).

Sample preparation and hybridization.

cRNA was generated by using the Affymetrix One-Cycle Target Labeling and Control Reagent kit (Affymetrix Inc., Santa Clara, California, USA), following the manufacturer's protocol. Total RNA was extracted from biological duplicate samples and analyzed using Affymetrix GeneChips (Mouse Genome 430 2.0 Array) to determine the global gene expression patterns. The Mouse Genome 430 2.0 Array contains more than 45000 probe sets including approximately over 34000 well-substantiated mouse genes. Chips were washed and scanned on the Affymetrix Complete GeneChip Instrument System and processed into CEL files.

Normalization

Data in the form of CEL files were background-subtracted and normalised with the Robust Multi-chip Average (RMA) pre-processing (Cope et al.,

2004; Irizarry et al., 2003b) that includes global background adjustment and quantile normalization. After we performed normalization for baseline to median based in its measured levels of expression across experiments. Moreover CEL files were normalised using Affymetrix MAS 5.0 algorithm to obtain the flags data and to use it to filter the RMA normalised data. For each probeset were averaged across biological replicates using the expression intensities to obtain the replicates-combined probeset intensity.

Note that, the normalization procedures are implemented in software GeneSpring GX 11.5 (http://www.chem.agilent.com/), all the arrays are used and no chip is discarded.

Probeset selection and filtering

A probeset selection algorithm was carried out to select a representative probeset for each gene, eliminating probesets that are annotated as cross-hybridizing to transcripts from different genes. A probeset was selected if and only if: A) It was not Absent in all samples; B) it possessed an Entrez GeneID in the Affymetrix probeset annotation database; C) its probeset name did not contain ‘‘_x_’’ or ‘‘_s_’’ and if it had a unique Entrez GeneID, otherwise it was a transcript of two distinct genes; D) it was not a hypothetical or predicted or Rik or CDNA clone or cDNA sequence gene.

This generated a list ~ 20000 probesets that constituted the genome-wide set. In ordered to remove genes that did not fluctuate with time in all samples which are often unexpressed/low expressed genes, only a $\log_2$ transformed fold change gene expression values of < 0.5, in at least one time point, are filter out. Moreover, in order to remove the replicate probes we referred to the values of T-test. The complete list of 9349 unique genes was used as working list.

Statistical and enrichment analysis

The differential expression of these genes over time was analyzed for statistical significance by one-way ANOVA analysis with unequal variance (Welch). P-values were calculated for each gene over the time course of hours 0, 24, 48 and 72 by every time-course. The calculations were performed on the $\log_2$-fold change in gene expression for time *n* versus time 0 for every time-course separately. To limit the detection of false positives, the p-values were adjusted by the Benjamini and Hochberg false-discovery-rate method with a cut-off of 0.001.

From the results list of 1210 genes, the functional GO and KEGG pathways were selected by taking into account of its gene expression levels. The KEGG pathways and Gene Ontology enrichment analysis were identified as significantly altered by using a hypergeometric test and p-value cut-off 0.05 and FDR $\leq$ 0.05. The results were displayed on Tables 1-3, and Figure 14.

Protein-protein interaction analysis was performed using all genes and the corresponded proteins were then loaded into GeneSpring GX 11.5. The protein-protein interaction (PPI) network was assessed using a variety of databases that include: published literature abstracts using a proprietary Natural Language Processing (NLP) algorithm, the experimentally reported physical interactions data parsed from IntAct (www.ebi.ac.uk/intact) that includes data from other databases like BIND (http://bond.unleashedinformatics.com/) and MINT (http://mint.bio.uniroma2.it/mint/Welcome.do). The PPI network was then visualized using GeneSpring software. Results are displayed on Figure 14B.

7.2 Data recovery and integration from public databases uncovers transformation-specific transcriptional downregulation of cAMP-PKA pathway-encoding genes.

Data recovery and normalization

Gene expression data of NCI60 cell lines and normal tissues samples were downloaded from the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/). In particular, gene expression profiles of NCI60 cell collection (cancer samples) were recovered from GEO database (GSE5949, (Wang et al., 2006)) in which the experimental data were obtained by using the Affymetrix HG-U95Av2 oligonucleotide array platform. For the analysis only results obtained by oligonucleotide arrays were considered, because this platform uses a different method to evaluate mRNA expression as compared to cDNA array platform. Therefore, also for normal tissue samples, the data used for the comparative analysis, were recovered from transcriptional profiles produced by using U95Av2 oligonucleotide array (GSE96 (Su et al., 2002), GSE6731 (Wu et al., 2007) and GSE1402 (Barnes et al., 2004b)).

A total of 81 transcriptional profiles encompassing cancer cell lines with nine histological origins and samples from six normal tissues were recovered. Further details can be found in the legends of Tables 4 and 5. All datasets were generated by downloading and processing CEL files. They were preprocessed using Robust Multichip Average (RMA) (Cope et al., 2004; Irizarry et al., 2003b) and then transformed from $\log_2$ values to linear scale values, and normalized per gene to the median value of its level of expression across 81 samples, as implemented in GeneSpring GX 7.3.1 (http://www.chem.agilent.com/). Note that, RMA implemented in GeneSpring GX 7.3.1, all the arrays are used and no chip is discarded.

Transformation-dependent, transcriptional remodeling of the PKA pathway-encoding genes in 60 human cancer cell lines (NCI60) and 21 human normal tissues

We identified and gathered the transcriptional profile for 41 genes encoding proteins involved in the PKA pathway (adenylyl cyclases -ADCY-, phosphodiesterases -PDE-, A-kinase anchor proteins -AKAP-, cAMP-dependent transcriptional factors -TF-, PKA catalytic subunits -PRKAC- and PKA regulatory subunits -PRKACR-, Table 6).

In order to identify specific variations in the expression pattern of the selected PKA pathway-related genes both in normal and transformed samples, different tools of analysis were used.

Initially, the PKA pathway related genes expression profiles, observed in transformed samples as compared to normal samples, were evaluated by analysis of variance (ANOVA). Such statistical linear modeling procedure, that partitions the total variance into parts corresponding to various sources in the model (Fisher, 1925) have been successfully used to analyze microarray data (Kerr et al., 2000; Pavlidis and Noble, 2001). In order to model and test the hypothesis that the expression of genes of PKA pathway was different between normal tissues and transformed cell lines, the following comparisons were used: Expression of gene *i* (where *i* is *i-esimo)* of Normal Tissues vs. Transformed cell lines (Figure 17), and a p-value < 0.05.

The same data-set was then analyzed through unsupervised hierarchical clustering (Johnson, 1967) (as implemented in the GeneSpring platform). Two-way hierarchical clustering was performed on RMA-generated linear scale expression levels using the Pearson correlation coefficient as the measure of similarity and complete linkage clustering (Eisen et al., 1998). The results of this process are dendrograms, in which short branches connect very similar elements, and longer branches join elements with diminishing

degrees of similarity. The vectors used were sample - normal tissues and transformed cells- and expression of genes of PKA pathway-related genes and the arms were classified by different variables: Conditions and Tissues, (Figure 18).

Analysis of mutational status of the NCI60 cell lines and correlation with tissue-specific PKA pathway gene regulation

The 60 cell lines were sorted according to mutational status, using the information provided by Catalogue Of Somatic Mutations In Cancer (http://www.sanger.ac.uk/genetics/CGP/cosmic/). This database holds somatic mutation data and other information related to human cancer cell lines and tissues, and can be interrogated through a series of web pages to provide a graphical or tabular view of the data along with various export options. We could sort the NCI60 cell lines in 4 subgroups presenting mutational activation of genes encoding components of the Ras pathway, of the PI3K pathway, of other pathways or for which no information was available, (Table 7).

In order to identify specific variations in the expression pattern of the PKA pathway-related genes in these 4 subgroups, different tools of analysis were used.

We applied unsupervised Principal Component Analysis (PCA) (Joliffe and Morgan, 1992) to establish the interrelationships among the samples used in our study. PCA on the mean centered and scaling data was used to model the effects of oncogene-dependent transformation on the gene expression. The following comparisons were performed: Expression of gene of Normal Tissues vs. PI3K mutation cell lines; vs. Ras mutation cell lines; vs. Not Tested mutation cell lines; vs. Other Mutation cell lines.

Also in this case, in order to model and test the hypothesis that the expression of genes of PKA pathway was different between normal tissues

and the four subgroups previously identified, we applied one-way ANOVA by using the following comparisons: Expression of gene *i* (where *i* is *i-esimo)* of Normal Tissues vs. PI3K mutation cell lines; vs. Ras mutation cell lines; vs. Not Tested mutation cell lines; vs. Other Mutation cell lines, (Figure 19D), and a p-value < 0.05.

The data-set of 41 genes was then analyzed through unsupervised hierarchical clustering (Pearson correlation coefficient and complete linkage clustering). The vectors used were sample - oncogene-dependent transformed cells - and expression of genes of PKA pathway-related genes. The results of this process are dendrograms, in which the arms were classified by different variables: Tissue, mutation and Pathway (Figure 20).

Computational analysis of promoters of differentially regulated PKA pathway-encoding genes and identification of transcriptional factor binding sites

In order to identify Transcriptional Factor Binding Sites (TFBS) present in promoters of co-regulated genes, the 41 PKA pathway-encoding genes were sorted, relative to their level of expression, in three groups: Strong (>1), Average (=1) and Low (<1), where 1 is the expression value calculated by RMA. Each groups was identified in each sample group, i.e., Normal Tissues, cell lines carrying mutation(s) in Ras pathway-encoding genes, cell lines carrying mutation(s) in PI3K pathway-encoding genes, cell lines carrying mutation(s) in other pathways, cell lines Not Tested for mutation, thus generating 15 subgroups. A TFBS was called present only when present in more than 70 % of promoters within each group.

Proximal promoter regions - defined as 500 nt upstream and 100 nt downstream from the transcription start site (TSS), automatically assigned to genes on the basis of 5' cap-site databases integrated into promoter identification program - were identified using Eldorado (GEMS launcher, Genomatix) and the Genomatix Promoter Database (Frech et al., 1997).

TFBS in the promoter regions were identified by using ModelInspector and the Genomatix Promoter Database, comprising a total of 519 matrices from 154 families (Matrix Family Library, on Vertebrates, Version 7.1, June 2008). The Matrix Family Library is based on 260000 human, mouse, and rat promoter sequences, with an average length of 650bp. Analysis on the 41 PKA pathway-encoding genes was performed with a threshold of 1.0 for the core similarity -that is reached only when the highest conserved bases of a matrix match exactly in the sequence- and a value of 0.85 for the Optimized matrix threshold (Cartharius et al., 2005). Optimized matrix threshold is the optimized value defined in a way that a minimum number of matches is found in non-regulatory test sequences. This value, when is higher than 0.80, permits the reduction of false positive matches.

The total number and frequency (i.e., the ratio between the total number of TBFS and the number of promoters present within each subgroup) of each TFBS within each subgroup were calculated. The frequency of each TFBS called present in each of the 15 subgroups of PKA pathway-encoding genes was compared with the frequency of the same TFBS within the Matrix Family Library on Vertebrates. TFBS enrichment was scored based on p-value generated by hypergeometric distribution and calculated with the 2-tailed Fisher's exact test, implemented through the use of a 2 x 2 contingency table (Figure 21).

In order to identify differences between the 15 groups a two-way hierarchical clustering (by using as vectors sample and TFBS) was applied by using the total number values and the frequency values of each TFBS identified in $\geq$ 70% of the promoters in each group. The total number value was transformed in the $\log_2$ and used in the hierarchical clustering by using the Pearson correlation coefficient as the measure of similarity and complete linkage clustering (Figure 22).

Promoter data mining

To identify known transcription factor binding sites in the promoter sequences of PKA pathway-encoding genes, the annotated promoter and associated information have been retrieved from Transcriptional Regulatory Element Database (TRED) (Zhao et al., 2005) and from NCBI (http://www.ncbi.nlm.nih.gov/). Both web sites are freely accessible. The results have been shown in the Table 9.

7.3 Comparative transcriptional analysis between a K-ras mouse cell model of transformation and the NCI60 human cancer cells collection.

Data recovery and normalization

Microarray datasets of human panel, composed of NCI60 cell lines (Ross et al., 2000; Scherf et al., 2000; Wang et al., 2006), immortalized cell lines and normal tissues, have been recovered from publicly accessible gene expression profile dataset (GEO) (Barrett et al., 2005; http://www.ncbi.nlm.nih.gov/geo/) or CellMiner database (http://discover.nci.nih.gov/cellminer/home.do). Microarray datasets of mouse panel, composed of MEF cells, immortalized and transformed cells, have been recovered from publicly accessible gene expression profile dataset (MEF) and data collected in our lab. Data collections are shown in Table 10. Mouse and human data were normalized separately. Data in the form of CEL files were background-subtracted and normalised with the Robust Multi-chip Average (RMA) pre-processing (Cope et al., 2004; Irizarry et al., 2003b) using the software GeneSpring GX 11 (Silicon Genetics) (http://www.chem.agilent.com/), RMA includes global background adjustment and quantile normalization (Allison et al., 2006; Irizarry et al., 2003b). Representation for baseline to median based in its measured levels of expression across experiments was performed. Moreover CEL files were

normalised using Affymetrix MAS 5.0 algorithm to obtain the flags data which were used to filter the RMA normalised data.

In this study, the samples have been classified in three ways: 1) Type: For human samples, primary normal tissues 'hNT', cancer 'hCCL' and immortalized 'hICL' cell lines; for mouse samples, Mouse Embryonic Fibroblasts 'mNT', cancer 'mCCL' and immortalized 'mICL' cell lines; 2) Tissue of origin, for human samples: Breast 'BR', CNS 'CNS', Colon 'CO', Lung 'LC', Leukemia 'LE', Melanoma 'ME', Ovary 'OV', Prostate 'PR' and Renal 'RE'; 3) Mutation: mutations able to interfere with the Ras pathway 'RAS', mutations able to interfere with PI3K-Akt pathway 'PI3K', no somatic mutations interfering with the two above pathways 'OTHER', and somatic mutations interfering with the above pathways has not been searched 'NOT TESTED' (Balestrieri et al., 2009).

Global analysis of human and mouse datasets

Principal component analysis (PCA) (Joliffe and Morgan, 1992; Raychaudhuri et al., 2000) was performed with GeneSpring. By visualizing projections of these components in low-dimensional spaces, it is possible to observe the grouping of samples, reflecting underlying patterns in their gene expression profiles. PCA on the mean centered and scaling data of all probe set on array was used and the Type classification of samples was performed. The percentage of temporal variance captured by the first four temporal PCs was 100% in both species. Results are shown on Figures 24 and 25.

Mouse-human gene orthologues: Probeset selection and Filtering

To extract orthologue identities, the ENTREZ GENE database (http://www.ncbi.nlm.nih.gov/gene) was queried using mouse or human identities provided in the Affymetrix annotation. The Affymetrix GeneChip platforms contain the following number of probe set: Mouse 430 2.0 , 45101 mouse array; U133A, 22283 human array. Between the mouse Mouse 430

2.0 and human U133A platforms, there can be found 43514 mouse-human orthologue pairs (18320 Human probe set and 23965 Mouse probe set). Affymetrix Web site (http://www.affymetrix.com) annotations for human HG-U133A and mouse MOUSE430_2 were downloaded from (http://www.affymetrix.com/products_services/arrays/specific/ht_hgu_133_ap.affx#1_4) and (http://www.affymetrix.com/products_services/arrays/specific/mouse430_2.affx#1_4)

The mouse and human data were analyzed separately. A probeset selection algorithm was carried out to select a representative probeset for each gene, eliminating probesets that are annotated as cross-hybridizing to transcripts from different genes. In particular a probeset was rejected if expression value are too close to background; they hybridize with transcripts of two or more distinct genes; and they do not have a orthologue pairs. After this selection 10478 and 13382 probesets were obtained, respectively by human and mouse databases. These two lists were overlapped and a list of 21606 mouse-human orthologue probe sets was obtained.

The first cross-species analysis

SOM Clustering (Kohonen, 1995; Vesanto, 1999; Vesanto and Alhoniemi, 2000) is based on a divisive approach where the input, entities and/or conditions are partitioned into a fixed user defined number of clusters. An entity and/or condition is assigned to a node (winning node), on this grid based on the similarity of its reference vector and the expression vector of the entity and/or condition. When an entity and/or condition is assigned to a node, the reference vector is adjusted to become more similar to the assigned entity and/or condition. The reference vectors of the neighboring nodes are also adjusted similarly, but to a lesser extent. This process is repeated iteratively to achieve convergence, where no entity and/or condition changes its winning node. Thus, entity and/or condition with similar expression

vectors get assigned to partitions that are physically closer on the grid, thereby producing a topology that preserves the mapping from input space onto the grid. The obtained proto-clusters (nodes in the grid) were clustered using hierarchical clustering, to produce a dendrogram based on the proximity of the reference vectors, so a combination of both hierarchical clustering and SOM was performed with GeneSpring and used to interpret the results. SOM and Hierarchical cluster was performed to mouse and human samples classified by Type. To fed the SOM algorithms, 21606 mouse-human orthologue probe sets were used. Result on Figure 26A. The chosen parameters were the followings: similarity measure: Pearson Cantered; max number of interaction 100; Grid Topology: Rectangular with rows 2 × columns 3; Initialization: 7 vectors; Neighborhood: Bubble; Learning Rate (α): 0.03. In Figure 26B the SOM analysis was performed with mouse and human samples, classified by Mutation. The chosen parameters were the followings: similarity measure: Pearson Absolute; max number of interaction 100; Grid Topology: Rectangular with rows 3 × columns 4; Initialization: 7 vectors; Neighborhood: Bubble; Learning Rate (α): 0.03.

Cross-species analysis by altered pathways and altered genes

The analysis of variance (ANOVA) was performed with GeneSpring. It is a statistical linear modeling procedure, that partitions the total variance into parts corresponding to various sources in the model (Fisher, 1993; Fisher, 1925) have been successfully used to analyze microarray data (Coombes et al., 2002; Kerr et al., 2000; Pavlidis and Noble, 2001; Pritchard et al., 2001). In order to model and test the hypothesis of a differential gene expression between samples, the following comparisons were used in both species, separately: *Expression of gene*$_i$ ~ CCL vs NT. ANOVA test was used and probe sets with a p-value ≤ 0.02 with FDR ≤ 0.1 and fold change ≥ +2 or ≤ -2 were selected. To select only one-to-one probe set for gene the p-values

obtained were used. In particular, the probe sets with the lower p-value for both species were selected. The fold change values of these genes were used to compare across species. Results are displayed on Table 11.

The pathway-specific analysis can be used to obtain a relatively comprehensive evaluation of the pathways to CCL-NT comparison across human and mouse species. KEGG pathways (http://www.genome.jp/kegg/pathway.html) whose use gene expression profile differed significantly in CCL-NT pair comparisons (resulting by ANOVA analysis) were identified using GeneCodis tools (http://genecodis.dacya.ucm.es/, ; Nogales-Cadenas et al., 2009). GeneCodis is a grid-based tool (web server application.) that integrates different sources of biological information to search for biological features (annotations) that frequently co-occur in a set of genes and rank them by statistical significance. Two different methods were used for identifying significantly altered pathways. First, were identified as significantly altered by performing a functional enrichment analysis on genes identified as significant by ANOVA analysis without other selection. The total list of genes has been given to feed the algorithm. A second method was applied to KEGG pathway genes in order to detect the shift of regulation of pathways, that were not detectable using a total genes. For this method, genes identified as significant by ANOVA analysis and their fold change values were used in order to obtained 2 different list: up regulated gene list and down regulated gene list. In this way, functional pathways were selected by taking into account of its gene expression levels. In the both ways the KEGG pathways were identified as significantly altered by using a hypergeometric test and p-value cut-off 0.05 and FDR $\leq$ 0.05. These analyses were performed for mouse and human genes, separately. The overlapping altered pathways are displayed on Tables 12, 13 and 14.

The selected common genes that were significantly changed in both species subsequently were classified in two groups in relation if they had a similar behavior in human and mouse comparisons. These two groups (96 and 403 genes) were used to compute enrichment on Gene Ontology terms, using as set of references the genome of human. Enrichment analysis was performed using Fisher's Exact Test that compares the observed frequency of each present GO term to the frequency in a reference set. The enrichment on GO was performed using the Anduril program and with p-value cut-off of 0.05. The program computes also Multiple Comparison Correction with different statistical tests. Visualization of enriched GO terms was created as networks for each GO ontology (see results on Figures 28 and 30).

Genes of Nutshell were combined with the corresponding proteins. PPI's analyses were performed in the first in order to observe a Biological Process rearrangements. Using all genes of Nutshell and the corresponded proteins were then loaded into GeneSpring GX 11 using the Human database. The protein-protein interaction (PPI) network was assessed using a variety of databases that include: published literature abstracts using a proprietary Natural Language Processing (NLP) algorithm, the experimentally reported physical interactions data parsed from IntAct (www.ebi.ac.uk/intact) that includes data from other databases like BIND (http://bond.unleashedinformatics.com/) ect. The PPI network was then visualized using GeneSpring software. The set of chosen parameters were the followings: Relations score $\geq$ 9; Relation Types: Biological Process; Entity local connectivity $\geq$ 2. Results are displayed on Figure 29.

Cross-species analysis from normal tissue to cancer cell lines, a nutshell of transformation

Hierarchical clustering is one of the simplest and widely used clustering techniques for analysis of gene expression data (Eisen et al., 1998; Johnson, 1967). The method follows an agglomerative approach, where the most

similar expression profiles are joined together to form a group. These are further joined in a tree structure, until all data forms a single group. The dendrogram is the most intuitive view of the results of this clustering method, in which short branches connect very similar elements, and longer branches join elements with diminishing degrees of similarity. The hierarchical clustering shown in the Figure 31 was performed on fold change value obtained from mCCL-mNT, mCCL-mICL and mICL-mNT comparison pairs, using the Euclidian distance matrix as the measure of similarity and centroid linkage clustering.

PPI's analysis were performed in order to identified a direct interaction *intra* and *inter* genes. The genes of Nut (27 genes) and the corresponded proteins were then loaded into **P**rotein **I**nteraction **N**etwork **A**nalysis (PINA) platform (http://csbi.ltdk.helsinki.fi/pina/, ; Wu et al., 2009). PINA is an integrated platform for protein interaction network construction, filtering, analysis, visualization and management. It integrates protein-protein interaction data from six public curate databases and builds a complete, non-redundant protein interaction dataset for six model organisms. Moreover, it provides a variety of built-in tools to filter and analyze the network for gaining insight into the network. At the same level of information analysis with 403 genes were performed using a GeneSpring database, in with more redundant information are filtered out. The resulting integration of these two approaches was a network of 119 nodes and 131 edges (see Figure 32A). From this network, sub-networks were extrapolated using the detects densely connected regions in large protein-protein interaction networks that represents the minimum pathway of interaction that link the principal hubs of the network. Results are displayed on Figure 32B.

7.4 Promoter Scan: Algorithm to detect over-represented TFBSs in the proximal promoter regions of co-regulated or co-classified genes.

All programs used in this paper are implemented into Anduril framework (Ovaska et al., 2010).

Data Collection

We retrieved three different data set collections of promoter regions: Test Set, Background and Random Set. In particular, every sequence containing 500 bp upstream and 199 bp downstream of the transcriptional starting point (TSS). This region of 700 bp represents the proximal promoter region. The gene chromosome location and its TSS are obtained from Ensemble database (http://www.ensembl.org/index.html) and the referent FASTA sequences is generated by EnsemblDNA component.

Test Set

This data set collection contain a total of 228 human promoter sequences as well the corresponding set of mouse homologues gene promoter regions. Each of these human genes promoter are characterized by exactly know of the capacity of at least one known regulatory element (TF) of bind the promoter region (for more detailed see Table 15).

Background and Random Set

The second data set of 1000 random human promoter sequences, are collected. For every of these genes, we identify a corresponding set of mouse homologues gene promoter regions. The total of human sequences and mouse sequences form sets called Background Set (2000 sequences). The promoter region genes in these sets are supposed to contain a normal distribution of TFs.

The third data set is called Random Set and is composed by 456 human and mouse sequences, taken random on their genome, as Background Set.

Transcriptional Factors collection

The transcription binding sites collection composed by 79 human TFs and their relative binding sites (TFBSs) were recovered from JASPAR database (http://jaspar.genereg.net/).

Matrix of TFBS match occurrences

All the promoter region sequences of different data set were submitted to MotifMatch component, that aligns the given motifs of transcriptional factor collection against the DNA sequences. The results were used as input of Promoter Scan algorithm.

Conclusions

As emphasized in the introduction, GeneChip technology allows the parallel and simultaneous detection of more than 30000 genes in cells. Although large genome-scale cDNA screens are powerful and efficient, they examine only one gene at a time, and will not uncover biological activities that often rely on multiple collaborating genes. Thus, GeneChip is one of the best high-throughput assay available for studying complex biological processes at a transcriptional level.

The GeneChip technology in cancer research is useful for the identification of information about disease-associated molecular signatures derived from analysis of the expression of basically all genes, as well as in the diagnostic decision. Therapeutic investigation targeting of genes and their regulatory mechanism may be used to complete existing therapies to halt the development and progression of cancer. Nevertheless the development of new analytical methods for DNA microarray data is a critical step to increase the sensibility of this technology.

The transition from the molecular level to the system level, promise to revolutionize our understanding of complex biological systems and provide new opportunities for practical application of such knowledge. The success of DNA technologies and the digital revolution brought about by the growth of the Internet have ensured that huge volumes of high-dimensional microarray expression data are now available. Data mining is an evolving and growing area of research and development. The problem is to mine useful information or patterns from the huge datasets. Microarrays provide a powerful basis to monitor the expression of tens of thousands of genes in order to identify mechanisms that govern the gene expression in an organism. The huge volume of such data, and their high dimensions, make gene expression data suitable candidates for the application of data mining functions. Therefore, it is expected that deeper computational integration of transcriptional data with other genome-wide findings, including -but not

limited to- proteomics, interactomics and metabolomics, will allow a better extraction of hidden information.

In this regard, in this thesis I used different statistical and computational approaches, typical resources applied in several scientific fields as mathematics, physics, bioinformatics, and I tried to develop new ideas to integrate and extract information that could brought us much closer to understanding the behavior of a cancer cell model on different conditions of investigation. I propose that such data integration can be further applied to examine the topology of biological networks, to provide information on directionality of interactions, and create wiring diagrams that better depict the functional outcome of component-component relationships. Together, these strategies should facilitate a systems approach to modular biology.

In the development of the current thesis I try to tackle this topic through four different works:

1) *Gene expression profiles comparative analysis of immortalized and K-Ras transformed mouse fibroblasts grown in different glucose availability;*
2) *Data recovery and integration from public databases uncovers transformation-specific transcriptional downregulation of cAMP-PKA pathway-encoding genes;*
3) *Comparative transcriptional analysis between a K-ras mouse cell model of transformation and the NCI60 human cancer cells collection;*
4) *Promoter Scan: Algorithm to detect over-represented TFBSs in the proximal promoter regions of co-regulated or co-classified genes.*

Since different biological conditions may have distinctive patterns of gene expressions, the identification of these patterns is the goal of microarray experiments. With this aim, in the first work of my thesis, I have

investigated the temporal effects of glucose shortage in normal and transformed mouse fibroblasts, by an high-throughput transcriptional analysis. This high-throughput gene expression analysis, conducted across the time by using several bioinformatic tools (hierarchical clustering, enrichment information analysis ect), it has allowed to elucidate the time-dependent role of the differentially expressed genes. In particular, our analysis permitted to identify some specific pathways (e.g. Metabolism, Cell Cycle, Signal Transduction, Apoptosis ect) that are strongly dependent both to k-ras oncogene and glucose availability.

During my thesis work, I have shown that gene expression analysis performed taking also in account gene function, interaction and regulation s is a helpful strategy to better understand where selective pressure is acting and the possible biological meaning it could have. A good example of this idea is represented by the second work of my thesis. Here I described a bioinformatics workflow able to show that those genes involved in the cAMP-PKA pathway are the ones under K-ras transformation pressure.

After several decades of cancer research, many details of the underlying mechanisms of cancer at the gene level and evolution conservation are still unclear. In the third work, in order to identify differential expressed genes between normal tissues/cells and cancer cells and common to mouse and human models, I propose an integrative method based on the direct comparison of a large number of microarray datasets generated from nine different types of human cancer cell lines (NCI60 cell collection) and a K-ras mouse model of transformation. The significant similarity identified between human and mouse models, both in terms of gene expression and pathway analysis, strongly support the use of cell/animal models to dissect human cancer. Importantly our analysis, permitted the identification of a restrict set of genes (Nutshell) of transformation that could be a key to open a potential door for further cancer research.

Working on scenarios as complete as possible (as metabolic networks, signaling cascades, biosynthetic pathways, transcriptional factor regulation, ect) is the best line of attack to understand the development and progression of cancer. In the fourth work, I have developed and described a new algorithm that has the ability to identify the most significant transcriptional factor binding sites in the promoters of a cluster of co-regulated genes. The algorithm shows a good degree of sensitivity and specificity and the results demonstrate its effectiveness to reduce false positive data. In the future the algorithm will be refined and compared with other methods having similar scope.

Altogether my results, obtained by bioinformatic approaches, demonstrate that the global delineation of complex cellular networks, will lead to a deeper knowledge of the complex cellular processes. This approach is the main core of the new field called Systems Biology. Although systems biology is in its infancy, it is already a vital part of modern biomedical research. Its potential benefits are enormous in both scientific and practical terms. Advances in the field will enable us to construct mechanistic models for the operation of the cellular systems, test and refine them using experimental approaches, and gradually witness the emergence of robust, dynamic, adapting, and developing systems from the information encoded in the genomes. Gaining such understanding will elucidate the causes of cancer development and survival, pointing the way to novel strategies for rational intervention in pathological conditions and the design of improved personalized drugs.

Acknowledgments

I would like to acknowledge and extend my sincere thanks to several people who greatly contributed to this work in various ways. Their expertise, experience, and support have made this work more valuable.

Dr. *Ferdinando Chiaradonna*, my supervisor for your support, enthusiasm, advice and patience. I am really grateful for your guidance throughout the years. I have always learnt from you and truly admire your brilliant knowledge and interconnection in the art of science. You have inspired me in a lot of ways, giving me answers but also questions.

I want to thank everyone in the Chiaradonna's research team: *Daniela Gaglio,* for always having an open door. I very much appreciate our scientific and non-scientific discussions. *Marco Gaviraghi*. I sincerely appreciate your great help with many things throughout in *n*-spaces and *n*-times. *Lara Sala Danna*. I appreciate your friendship, help and generosity. You are incredible. *Roberta Palorini*, for all the help, advice and being so friendly in special with the rules of crossword. *Andrea Monestiroli*, for long coffee breaks.

I would also like to express my gratefulness to Prof. Lilia Alberghina and Prof. Marco Vanoni, for giving me the opportunity to conduct an interesting research.

I would also like to express my sincere thanks to Prof. *Sampsa Hautaniemi*, for giving me the opportunity to work in Finland. In addition to the professional advising, his encouragement and support have played an essential part in accomplishing the results leading to completion of my thesis. Many thanks to *Hautaniemi's laboratory*: Tiia, Erkka, Jianmin, Lilli, Chengyu, Viljami, Kari, Riku, Ali, Ville, Kristian, Anna-Maria, Minna, Marko, Sirkku, Ping, Vladimir and Javier. These people have done great work in teaching me in computational programming and they made possible the wonderful Finnish experience.

Thanks to my *old* and *new friends* for taking my mind when needed with interesting and passionate discussions: Lavinia, Silvia, Ambrogio, Pasquale, Filippo, Tiziana, Alessio, Raffaele, Cristina, Gloria, Alessandro, Davide, Alexander, Clo, Boris, Temesghen, Pietro, Heli, Samik, Emanuele, Kalina, Daniele, Flavio, Fabrizia, William, Alexandra, Carmelo, Grazia, Daniele, Michela, Laura, and many others.

I would also like to express my sincere thanks to my mother *Lucia* and my dearest sister *Federica* for your patience, encouragement and support. Thank you for always making me feel so special.

References

Acin-Perez, R., Salazar, E., Kamenetsky, M., Buck, J., Levin, L. R., and Manfredi, G. (2009). Cyclic AMP produced inside mitochondria regulates oxidative phosphorylation. Cell Metab *9*, 265-276.

Affymetrix (2002). Affymetrix: Microarray Suite 5.0 User's Guide., (Santa Clara, CA, USA: Affymetrix).

Alberghina, L., Chiaradonna, F., and Vanoni, M. (2004). Systems biology and the molecular circuits of cancer. Chembiochem *5*, 1322-1333.

Alchanati, I., Nallar, S. C., Sun, P., Gao, L., Hu, J., Stein, A., Yakirevich, E., Konforty, D., Alroy, I., Zhao, X.*, et al.* (2006). A proteomic analysis reveals the loss of expression of the cell death regulatory gene GRIM-19 in human renal cell carcinomas. Oncogene *25*, 7138-7147.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. Nature reviews *7*, 55-65.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America *96*, 6745-6750.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences of the United States of America *97*, 10101-10106.

Altman, R. B., and Raychaudhuri, S. (2001). Whole-genome expression analysis: challenges beyond clustering. Curr Opin Struct Biol *11*, 340-347.

Anderson, R. A., Boronenkov, I. V., Doughman, S. D., Kunz, J., and Loijens, J. C. (1999). Phosphatidylinositol phosphate kinases, a multifaceted family of signaling enzymes. J Biol Chem *274*, 9907-9910.

Augenlicht, L. H., Kobrin, D., Pavlovec, A., and Royston, M. E. (1984). Elevated expression of an endogenous retroviral long terminal repeat in a mouse colon tumor. J Biol Chem *259*, 1842-1847.

Augenlicht, L. H., Wahrman, M. Z., Halsey, H., Anderson, L., Taylor, J., and Lipkin, M. (1987). Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. Cancer Res *47*, 6017-6021.

Bader, A. G., Kang, S., Zhao, L., and Vogt, P. K. (2005). Oncogenic PI3K deregulates transcription and translation. Nat Rev Cancer *5*, 921-929.

Bagnato, A., Cirilli, A., Salani, D., Simeone, P., Muller, A., Nicotra, M. R., Natali, P. G., and Venuti, A. (2002). Growth inhibition of cervix carcinoma cells in vivo by endothelin A receptor blockade. Cancer Res *62*, 6381-6384.

Bagnato, A., and Spinella, F. (2003). Emerging role of endothelin-1 in tumor angiogenesis. Trends in endocrinology and metabolism: TEM *14*, 44-50.

Bailey, T. L., and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol *3*, 21-29.

Balestrieri, C., Alberghina, L., Vanoni, M., and Chiaradonna, F. (2009). Data recovery and integration from public databases uncovers transformation-specific transcriptional downregulation of cAMP-PKA pathway-encoding genes. BMC bioinformatics *10 Suppl 12*, S1.

Balkwill, F., and Mantovani, A. (2001). Inflammation and cancer: back to Virchow? Lancet *357*, 539-545.

Baracca, A., Chiaradonna, F., Sgarbi, G., Solaini, G., Alberghina, L., and Lenaz, G. (2010). Mitochondrial Complex I decrease is responsible for bioenergetic dysfunction in K-ras transformed cells. Biochim Biophys Acta *1797*, 314-323.

Barnes, G. L., Hebert, K. E., Kamal, M., Javed, A., Einhorn, T. A., Lian, J. B., Stein, G. S., and Gerstenfeld, L. C. (2004a). Fidelity of Runx2 activity in breast cancer cells is required for the generation of metastases-associated osteolytic disease. Cancer Res *64*, 4506-4513.

Barnes, G. L., Javed, A., Waller, S. M., Kamal, M. H., Hebert, K. E., Hassan, M. Q., Bellahcene, A., Van Wijnen, A. J., Young, M. F., Lian, J. B.*, et al.* (2003). Osteoblast-related transcription factors Runx2 (Cbfa1/AML3) and MSX2 mediate the expression of bone sialoprotein in human metastatic breast cancer cells. Cancer Res *63*, 2631-2637.

Barnes, M. G., Aronow, B. J., Luyrink, L. K., Moroldo, M. B., Pavlidis, P., Passo, M. H., Grom, A. A., Hirsch, R., Giannini, E. H., Colbert, R. A.*, et al.* (2004b). Gene expression in juvenile arthritis and spondyloarthropathy: pro-angiogenic ELR+ chemokine genes relate to course of arthritis. Rheumatology (Oxford, England) *43*, 973-979.

Barradeau, S., Imaizumi-Scherrer, T., Weiss, M. C., and Faust, D. M. (2001). Muscle-regulated expression and determinants for neuromuscular junctional localization of the mouse RIalpha regulatory subunit of cAMP-dependent protein kinase. Proceedings of the National Academy of Sciences of the United States of America *98*, 5037-5042.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res *33*, D562-566.

Basilevsky, A. (1994). Statistical Factor Analysis and Related Methods: Theory and Applications, (New York: John Wiley & Sons).

Beckner, S. K., Hattori, S., and Shih, T. Y. (1985). The ras oncogene product p21 is not a regulatory component of adenylate cyclase. Nature *317*, 71-72.

Bedell, M. A., Jenkins, N. A., and Copeland, N. G. (1997a). Mouse models of human disease. Part I: techniques and resources for genetic analysis in mice. Genes Dev *11*, 1-10.

Bedell, M. A., Largaespada, D. A., Jenkins, N. A., and Copeland, N. G. (1997b). Mouse models of human disease. Part II: recent progress and future directions. Genes Dev *11*, 11-43.

Bellis, A., Castaldo, D., Trimarco, V., Monti, M. G., Chivasso, P., Sadoshima, J., Trimarco, B., and Morisco, C. (2009). Cross-talk between PKA and Akt protects endothelial cells from apoptosis in the late ischemic preconditioning. Arterioscler Thromb Vasc Biol *29*, 1207-1212.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. J Comput Biol *6*, 281-297.

Bergmann, S., Ihmels, J., and Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. PLoS Biol *2*, E9.

Black, A. R., Black, J. D., and Azizkhan-Clifford, J. (2001). Sp1 and kruppel-like factor family of transcription factors in cell growth regulation and cancer. J Cell Physiol *188*, 143-160.

Blanchette, M., and Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res *12*, 739-748.

Bolstad, A. I., and Jonsson, R. (2002). Genetic aspects of Sjogren's syndrome. Arthritis Res *4*, 353-359.
Bolstad, B. M., Collin, F., Simpson, K. M., Irizarry, R. A., and Speed, T. P. (2004). Experimental design and low-level analysis of microarray data. Int Rev Neurobiol *60*, 25-58.
Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England) *19*, 185-193.
Boneau, C. A. (1960). The effects of violations of assumptions underlying the test. Psychological bulletin *57*, 49-64.
Bossu, P., Vanoni, M., Wanke, V., Cesaroni, M. P., Tropea, F., Melillo, G., Asti, C., Porzio, S., Ruggiero, P., Di Cioccio, V.*, et al.* (2000). A dominant negative RAS-specific guanine nucleotide exchange factor reverses neoplastic phenotype in K-ras transformed mouse fibroblasts. Oncogene *19*, 2147-2154.
Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. Genome Res *8*, 1202-1215.
Canales, R. D., Luo, Y., Willey, J. C., Austermiller, B., Barbacioru, C. C., Boysen, C., Hunkapiller, K., Jensen, R. V., Knight, C. R., Lee, K. Y.*, et al.* (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. Nature biotechnology *24*, 1115-1122.
Carpenter, G., and Cohen, S. (1990). Epidermal growth factor. J Biol Chem *265*, 7709-7712.
Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics (Oxford, England) *21*, 2933-2942.
Cervigne, N. K., Reis, P. P., Machado, J., Sadikovic, B., Bradley, G., Galloni, N. N., Pintilie, M., Jurisica, I., Perez-Ordonez, B., Gilbert, R.*, et al.* (2009). Identification of a microRNA signature associated with progression of leukoplakia to oral carcinoma. Human molecular genetics *18*, 4818-4829.
Chang, C. R., and Blackstone, C. (2007). Cyclic AMP-dependent protein kinase phosphorylation of Drp1 regulates its GTPase activity and mitochondrial morphology. J Biol Chem *282*, 21583-21587.
Chang, F., Steelman, L. S., Lee, J. T., Shelton, J. G., Navolanic, P. M., Blalock, W. L., Franklin, R. A., and McCubrey, J. A. (2003). Signal transduction mediated by the Ras/Raf/MEK/ERK pathway from cytokine receptors to transcription factors: potential targeting for therapeutic intervention. Leukemia *17*, 1263-1293.
Chao, J. R., Ni, Y. G., Bolanos, C. A., Rahman, Z., DiLeone, R. J., and Nestler, E. J. (2002). Characterization of the mouse adenylyl cyclase type VIII gene promoter: regulation by cAMP and CREB. The European journal of neuroscience *16*, 1284-1294.
Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. (1996). Accessing genetic information with high-density DNA arrays. Science (New York, NY *274*, 610-614.
Chen, X., Cheung, S. T., So, S., Fan, S. T., Barry, C., Higgins, J., Lai, K. M., Ji, J., Dudoit, S., Ng, I. O.*, et al.* (2002). Gene expression patterns in human liver cancers. Mol Biol Cell *13*, 1929-1939.
Chesler, E. J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H. C., Mountz, J. D., Baldwin, N. E., Langston, M. A.*, et al.* (2005). Complex trait analysis of gene

expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nature genetics *37*, 233-242.
Chiaradonna, F., Balestrieri, C., Gaglio, D., and Vanoni, M. (2008). RAS and PKA pathways in cancer: new insight from transcriptional analysis. Front Biosci *13*, 5257-5278.
Chiaradonna, F., Gaglio, D., Vanoni, M., and Alberghina, L. (2006a). Expression of transforming K-Ras oncogene affects mitochondrial function and morphology in mouse fibroblasts. Biochim Biophys Acta *1757*, 1338-1356.
Chiaradonna, F., Sacco, E., Manzoni, R., Giorgio, M., Vanoni, M., and Alberghina, L. (2006b). Ras-dependent carbon metabolism and transformation in mouse fibroblasts. Oncogene *25*, 5391-5404.
Cho-Chung, Y. S. (1990). Role of cyclic AMP receptor proteins in growth, differentiation, and suppression of malignancy: new approaches to therapy. Cancer Res *50*, 7093-7100.
Cho, R. J., Huang, M., Campbell, M. J., Dong, H., Steinmetz, L., Sapinoso, L., Hampton, G., Elledge, S. J., Davis, R. W., and Lockhart, D. J. (2001). Transcriptional regulation and function during the human cell cycle. Nature genetics *27*, 48-54.
Cho, S. Y., and Klemke, R. L. (2000). Extracellular-regulated kinase activation and CAS/Crk coupling regulate cell migration and suppress apoptosis during invasion of the extracellular matrix. J Cell Biol *149*, 223-236.
Chung, C. H., Bernard, P. S., and Perou, C. M. (2002). Molecular portraits and the family tree of cancer. Nature genetics *32 Suppl*, 533-540.
Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. Nature genetics *32 Suppl*, 490-495.
Clark, E. S., Whigham, A. S., Yarbrough, W. G., and Weaver, A. M. (2007). Cortactin is an essential regulator of matrix metalloproteinase secretion and extracellular matrix degradation in invadopodia. Cancer Res *67*, 4227-4235.
Claverie, J. M. (1999). Computational methods for the identification of differential and coordinated gene expression. Human molecular genetics *8*, 1821-1832.
Clegg, C. H., Haugen, H. S., and Boring, L. F. (1996). Promoter sequences in the RI beta subunit gene of cAMP-dependent protein kinase required for transgene expression in mouse brain. J Biol Chem *271*, 1638-1644.
Clegg, C. H., Koeiman, N. R., Jenkins, N. A., Gilbert, D. J., Copeland, N. G., and Neubauer, M. G. (1994). Structural features of the murine gene encoding the RI beta subunit of cAMP-dependent protein kinase. Molecular and cellular neurosciences *5*, 153-164.
Coombes, K. R., Highsmith, W. E., Krogmann, T. A., Baggerly, K. A., Stivers, D. N., and Abruzzo, L. V. (2002). Identifying and quantifying sources of variation in microarray data using high-density cDNA membrane arrays. J Comput Biol *9*, 655-669.
Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., and Speed, T. P. (2004). A benchmark for Affymetrix GeneChip expression measures. Bioinformatics (Oxford, England) *20*, 323-331.
Corden, L. D., and McLean, W. H. (1996). Human keratin diseases: hereditary fragility of specific epithelial tissues. Experimental dermatology *5*, 297-307.
Craig, J. C., Eberwine, J. H., Calvin, J. A., Wlodarczyk, B., Bennett, G. D., and Finnell, R. H. (1997). Developmental expression of morphoregulatory genes in the

mouse embryo: an analytical approach using a novel technology. Biochemical and molecular medicine *60*, 81-91.
Cribbs, J. T., and Strack, S. (2007). Reversible phosphorylation of Drp1 by cyclic AMP-dependent protein kinase and calcineurin regulates mitochondrial fission and cell death. EMBO Rep *8*, 939-944.
Cuevas, B. D., Winter-Vann, A. M., Johnson, N. L., and Johnson, G. L. (2006). MEKK1 controls matrix degradation and tumor cell dissemination during metastasis of polyoma middle-T driven mammary cancer. Oncogene *25*, 4998-5010.
D'Haeseleer, P. (2005). How does gene expression clustering work? Nature biotechnology *23*, 1499-1501.
D'Sa, C., Tolbert, L. M., Conti, M., and Duman, R. S. (2002). Regulation of cAMP-specific phosphodiesterases type 4B and 4D (PDE4) splice variants by cAMP signaling in primary cortical neurons. Journal of neurochemistry *81*, 745-757.
Dahle, M. K., Gronning, L. M., Cederberg, A., Blomhoff, H. K., Miura, N., Enerback, S., Tasken, K. A., and Tasken, K. (2002a). Mechanisms of FOXC2- and FOXD1-mediated regulation of the RI alpha subunit of cAMP-dependent protein kinase include release of transcriptional repression and activation by protein kinase B alpha and cAMP. J Biol Chem *277*, 22902-22908.
Dahle, M. K., Knutsen, H. K., Tasken, K. A., Pilz, R., and Tasken, K. (2001). Cyclic AMP regulates expression of the RI alpha subunit of cAMP-dependent protein kinase through an alternatively spliced 5' UTR. Eur J Biochem *268*, 5920-5929.
Dahle, M. K., Tasken, K., and Tasken, K. A. (2002b). USF2 inhibits C/EBP-mediated transcriptional regulation of the RIIbeta subunit of cAMP-dependent protein kinase. BMC molecular biology *3*, 10.
Dang, C. V., and Semenza, G. L. (1999). Oncogenic alterations of metabolism. Trends in biochemical sciences *24*, 68-72.
Dave, S. S., Wright, G., Tan, B., Rosenwald, A., Gascoyne, R. D., Chan, W. C., Fisher, R. I., Braziel, R. M., Rimsza, L. M., Grogan, T. M., *et al.* (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. N Engl J Med *351*, 2159-2169.
Davis, R. J. (1995). Transcriptional regulation by MAP kinases. Molecular reproduction and development *42*, 459-467.
Davis, R. J. (2000). Signal transduction by the JNK group of MAP kinases. Cell *103*, 239-252.
De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., and Moreau, Y. (2002). Adaptive quality-based clustering of gene expression profiles. Bioinformatics *18*, 735-746.
DeBerardinis, R. J., Lum, J. J., Hatzivassiliou, G., and Thompson, C. B. (2008). The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. Cell metabolism *7*, 11-20.
Dehmer, M., Borgert, S., and Emmert-Streib, F. (2008). Entropy bounds for hierarchical molecular networks. PLoS One *3*, e3079.
Dehmer, M., and Emmert-Streib, F. (2008). Structural information content of networks: graph entropy based on local vertex functionals. Comput Biol Chem *32*, 131-138.
Delfino, F. J., and Walker, W. H. (1999). NF-kappaB induces cAMP-response element-binding protein gene transcription in sertoli cells. J Biol Chem *274*, 35607-35613.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nature genetics *14*, 457-460.
Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. Nature *412*, 822-826.
Do, J. H., and Choi, D. K. (2006). Normalization of microarray data: single-labeled and dual-labeled arrays. Molecules and cells *22*, 254-261.
Downward, J. (2003). Targeting RAS signaling pathways in cancer therapy. Nat Rev Cancer *3*, 11-22.
Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., and Krawetz, S. A. (2003). Global functional profiling of gene expression. Genomics *81*, 98-104.
Duda, R. O., Hart, P. E., and Stork, D. G. (2001). Pattern Classification, (New York: John Wiley & Sons).
Dumaz, N., and Marais, R. (2005). Integrating signals between cAMP and the RAS/RAF/MEK/ERK signaling pathways. Based on the anniversary prize of the Gesellschaft fur Biochemie und Molekularbiologie Lecture delivered on 5 July 2003 at the Special FEBS Meeting in Brussels. The FEBS journal *272*, 3491-3504.
Dupuy, A. J., Morgan, K., von Lintig, F. C., Shen, H., Acar, H., Hasz, D. E., Jenkins, N. A., Copeland, N. G., Boss, G. R., and Largaespada, D. A. (2001). Activation of the Rap1 guanine nucleotide exchange gene, CalDAG-GEF I, in BXH-2 murine myeloid leukemia. J Biol Chem *276*, 11804-11811.
Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. (2002). A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics (Oxford, England) *18 Suppl 1*, S105-110.
Dyrskjot, L., Thykjaer, T., Kruhoffer, M., Jensen, J. L., Marcussen, N., Hamilton-Dutoit, S., Wolf, H., and Orntoft, T. F. (2003). Identifying distinct classes of bladder carcinoma using microarrays. Nature genetics *33*, 90-96.
Eisen, M. B., and Brown, P. O. (1999). DNA arrays for analysis of gene expression. Methods in enzymology *303*, 179-205.
Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*, 14863-14868.
Ellwood-Yen, K., Graeber, T. G., Wongvipat, J., Iruela-Arispe, M. L., Zhang, J., Matusik, R., Thomas, G. V., and Sawyers, C. L. (2003). Myc-driven murine prostate cancer shares molecular features with human prostate tumors. Cancer cell *4*, 223-238.
Faris, M., Latinis, K. M., Kempiak, S. J., Koretzky, G. A., and Nel, A. (1998). Stress-induced Fas ligand expression in T cells is mediated through a MEK kinase 1-regulated response element in the Fas ligand promoter. Molecular and cellular biology *18*, 5414-5424.
Fawal, M., Armstrong, F., Ollier, S., Dupont, H., Touriol, C., Monsarrat, B., Delsol, G., Payrastre, B., and Morello, D. (2006). A "liaison dangereuse" between AUF1/hnRNPD and the oncogenic tyrosine kinase NPM-ALK. Blood *108*, 2780-2788.
Fernandez, P. C., Frank, S. R., Wang, L., Schroeder, M., Liu, S., Greene, J., Cocito, A., and Amati, B. (2003). Genomic targets of the human c-Myc protein. Genes Dev *17*, 1115-1129.

Finco, T. S., and Baldwin, A. S., Jr. (1993). Kappa B site-dependent induction of gene expression by diverse inducers of nuclear factor kappa B requires Raf-1. J Biol Chem *268*, 17676-17679.
Fisher, L. (1993). Biostatistics: a methodology for the health sciences., (New York: John Wiley & Sons Inc.).
Fisher, R. A. (1925). Statistical Methods for Research Workers, (Edinburg: Oliver and Boyd).
Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. Science (New York, NY *251*, 767-773.
Foulds, C. E., Nelson, M. L., Blaszczak, A. G., and Graves, B. J. (2004). Ras/mitogen-activated protein kinase signaling activates Ets-1 and Ets-2 by CBP/p300 recruitment. Molecular and cellular biology *24*, 10954-10964.
Franks, D. J., Whitfield, J. F., and Durkin, J. P. (1987). Viral p21 Ki-RAS protein: a potent intracellular mitogen that stimulates adenylate cyclase activity in early G1 phase of cultured rat cells. Journal of cellular biochemistry *33*, 87-94.
Frech, K., Danescu-Mayer, J., and Werner, T. (1997). A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J Mol Biol *270*, 674-687.
Friedl, P., and Wolf, K. (2003). Tumour-cell invasion and migration: diversity and escape mechanisms. Nat Rev Cancer *3*, 362-374.
Gaglio, D., Soldati, C., Vanoni, M., Alberghina, L., and Chiaradonna, F. (2009). Glutamine deprivation induces abortive s-phase rescued by deoxyribonucleotides in k-ras transformed fibroblasts. PLoS One *4*, e4715.
Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I.*, et al.* (2001). Diversity of gene expression in adenocarcinoma of the lung. Proceedings of the National Academy of Sciences of the United States of America *98*, 13784-13789.
Garcia-Escudero, R., Martinez-Cruz, A. B., Santos, M., Lorz, C., Segrelles, C., Garaulet, G., Saiz-Ladera, C., Costa, C., Buitrago-Perez, A., Duenas, M., and Paramio, J. M. (2010). Gene expression profiling of mouse p53-deficient epidermal carcinoma defines molecular determinants of human cancer malignancy. Molecular cancer *9*, 193.
Gatenby, R. A., and Gillies, R. J. (2004). Why do cancers have high aerobic glycolysis? Nat Rev Cancer *4*, 891-899.
Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics (Oxford, England) *20*, 307-315.
Geho, D. H., Bandle, R. W., Clair, T., and Liotta, L. A. (2005). Physiological mechanisms of tumor-cell invasion and migration. Physiology (Bethesda) *20*, 194-200.
Geller, S. C., Gregg, J. P., Hagerman, P., and Rocke, D. M. (2003). Transformation and normalization of oligonucleotide microarray data. Bioinformatics (Oxford, England) *19*, 1817-1823.
Gelman, A. (2005). Analysis of variance—why it is more important than ever. Annals Statistics *33*, 1-53.

Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. Proceedings of the National Academy of Sciences of the United States of America *97*, 12079-12084.
Gibbons, F. D., and Roth, F. P. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. Genome research *12*, 1574-1581.
Gibson, G. (2003). Microarray analysis: genome-scale hypothesis scanning. PLoS Biol *1*, E15.
Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A.*, et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (New York, NY *286*, 531-537.
Gordon, A. D. (1999). Classification (2nd ed.), (London: Chapman Hall/CRC).
Gouble, A., Grazide, S., Meggetto, F., Mercier, P., Delsol, G., and Morello, D. (2002). A new player in oncogenesis: AUF1/hnRNPD overexpression leads to tumorigenesis in transgenic mice. Cancer Res *62*, 1489-1495.
Gruvberger-Saal, S. K., Cunliffe, H. E., Carr, K. M., and Hedenfalk, I. A. (2006). Microarrays in breast cancer research and clinical practice--the future lies ahead. Endocrine-related cancer *13*, 1017-1031.
Guillemin, M. C., Raffoux, E., Vitoux, D., Kogan, S., Soilihi, H., Lallemand-Breitenbach, V., Zhu, J., Janin, A., Daniel, M. T., Gourmel, B.*, et al.* (2002). In vivo activation of cAMP signaling induces growth arrest and differentiation in acute promyelocytic leukemia. J Exp Med *196*, 1373-1380.
Hall, C. L., Kang, S., MacDougald, O. A., and Keller, E. T. (2006). Role of Wnts in prostate cancer bone metastases. Journal of cellular biochemistry *97*, 661-672.
Hanash, S. (2004). Integrated global profiling of cancer. Nat Rev Cancer *4*, 638-644.
Harada, H., Becknell, B., Wilm, M., Mann, M., Huang, L. J., Taylor, S. S., Scott, J. D., and Korsmeyer, S. J. (1999). Phosphorylation and inactivation of BAD by mitochondria-anchored protein kinase A. Mol Cell *3*, 413-422.
Hardiman, G. (2004). Microarray platforms--comparisons and contrasts. Pharmacogenomics *5*, 487-502.
Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome biology *1*, RESEARCH0003.
Hazzalin, C. A., and Mahadevan, L. C. (2002). MAPK-regulated transcription: a continuously variable gene switch? Nat Rev Mol Cell Biol *3*, 30-40.
Herrero, J., Valencia, A., and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics (Oxford, England) *17*, 126-136.
Hertz, G. Z., and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics *15*, 563-577.
Herzig, M., and Christofori, G. (2002). Recent advances in cancer research: mouse models of tumorigenesis. Biochim Biophys Acta *1602*, 97-113.
Heyer, L. J., Kruglyak, S., and Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. Genome research *9*, 1106-1115.
Hillegass, J. M., Murphy, K. A., Villano, C. M., and White, L. A. (2006). The impact of aryl hydrocarbon receptor signaling on matrix metabolism: implications for development and disease. Biological chemistry *387*, 1159-1173.

Hilsenbeck, S. G., Friedrichs, W. E., Schiff, R., O'Connell, P., Hansen, R. K., Osborne, C. K., and Fuqua, S. A. (1999). Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. Journal of the National Cancer Institute *91*, 453-459.
Hoaglin DC, M. F., and Tukey JW. (1983). Understanding Robust and Exploratory Data Analysis, (New York: Wiley).
Hoaglin, D. C., Mosteller, F., and Tukey, J. (1983). Understanding Robust and Exploratory Data Analysis, (New York: Wiley).
Hodgson, G., Hager, J. H., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D. G., Pinkel, D., Collins, C., *et al.* (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. Nature genetics *29*, 459-464.
Hood, L., Heath, J. R., Phelps, M. E., and Lin, B. (2004). Systems biology and new technologies enable predictive and preventative medicine. Science (New York, NY *306*, 640-643.
Houslay, M. D., and Baillie, G. S. (2003). The role of ERK2 docking and phosphorylation of PDE4 cAMP phosphodiesterase isoforms in mediating cross-talk between the cAMP and ERK signaling pathways. Biochem Soc Trans *31*, 1186-1190.
http://arep.med.harvard.edu/mrnadata/mrnasoft.html.
http://bayesweb.wadsworth.org/gibbs/gibbs.html.
http://bond.unleashedinformatics.com/.
http://csbi.ltdk.helsinki.fi/pina/.
http://discover.nci.nih.gov/cellminer/home.do.
http://genecodis.dacya.ucm.es/.
http://jaspar.genereg.net/.
http://meme.sdsc.edu/.
http://mint.bio.uniroma2.it/mint/Welcome.do.
http://www.affymetrix.com/index.affx.
http://www.affymetrix.com/products_services/arrays/specific/ht_hgu_133_ap.affx14
http://www.affymetrix.com/products_services/arrays/specific/mouse430_2.affx#1_4
http://www.chem.agilent.com/
http://www.ensembl.org/index.html.
http://www.genome.jp/kegg/pathway.html
http://www.ncbi.nlm.nih.gov/
http://www.ncbi.nlm.nih.gov/geo/
http://www.sanger.ac.uk/genetics/CGP/cosmic/.
Hubbell, E. (2001). Multiplex sequencing by hybridization. J Comput Biol *8*, 141-149.
Hubbell, E., Liu, W. M., and Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics (Oxford, England) *18*, 1585-1592.
Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol *296*, 1205-1214.
Hyvarinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. Neural Netw *13*, 411-430.
Impey, S., McCorkle, S. R., Cha-Molstad, H., Dwyer, J. M., Yochum, G. S., Boss, J. M., McWeeney, S., Dunn, J. J., Mandel, G., and Goodman, R. H. (2004). Defining

the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. Cell *119*, 1041-1054.
Inman, C. K., and Shore, P. (2003). The osteoblast transcription factor Runx2 is expressed in mammary epithelial cells and mediates osteopontin expression. J Biol Chem *278*, 48684-48689.
Irgon, J., Huang, C. C., Zhang, Y., Talantov, D., Bhanot, G., and Szalma, S. (2010). Robust multi-tissue gene panel for cancer detection. BMC cancer *10*, 319.
Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res *31*, e15.
Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249-264.
Isidoro, A., Martinez, M., Fernandez, P. L., Ortega, A. D., Santamaria, G., Chamorro, M., Reed, J. C., and Cuezva, J. M. (2004). Alteration of the bioenergetic phenotype of mitochondria is a hallmark of breast, gastric, lung and oesophageal cancer. Biochem J *378*, 17-20.
Jain, A. K., and Dubes, R. C. (1988). Algorithms for Clustering Data, (Upper Saddle River, NJ: Prentice-Hall).
Jezek, P., Plecita-Hlavata, L., Smolkova, K., and Rossignol, R. (2010). Distinctions and similarities of cell bioenergetics and the role of mitochondria in hypoxia, cancer, and embryonic development. The international journal of biochemistry & cell biology *42*, 604-622.
Jiang, D. (2003). DHC: a density hierarchical clustering methods for time series gene expression data. In Third IEEE Symposium on BioInformatics and BioEngineering.
Johnson, R. A. (1998). Applied multivariate statistical analysis: Prentice Hall).
Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika *32*, 241-254.
Joliffe, I. T., and Morgan, B. J. (1992). Principal component analysis and exploratory factor analysis. Statistical methods in medical research *1*, 69-95.
Jordan, I. K., Marino-Ramirez, L., and Koonin, E. V. (2005). Evolutionary significance of gene expression divergence. Gene *345*, 119-126.
Jorissen, R. N., Walker, F., Pouliot, N., Garrett, T. P., Ward, C. W., and Burgess, A. W. (2003). Epidermal growth factor receptor: mechanisms of activation and signaling. Exp Cell Res *284*, 31-53.
Kahn, S., Yamamoto, F., Almoguera, C., Winter, E., Forrester, K., Jordano, J., and Perucho, M. (1987). The c-K-ras gene and human cancer (review). Anticancer research *7*, 639-652.
Karandikar, M., Xu, S., and Cobb, M. H. (2000). MEKK1 binds raf-1 and the ERK2 cascade components. J Biol Chem *275*, 40120-40127.
Kaufman, L. P., and Rousseeuw, J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis, (Hoboken, NJ: John Wiley & Sons).
Kedzierski, R. M., and Yanagisawa, M. (2001). Endothelin system: the double-edged sword in health and disease. Annual review of pharmacology and toxicology *41*, 851-876.
Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. J Comput Biol *7*, 819-837.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W., and Paabo, S. (2004). A neutral model of transcriptome evolution. PLoS Biol *2*, E132.
Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine *7*, 673-679.
Kim, J. H., Kim, H., Lee, K. Y., Kang, J. W., Lee, K. H., Park, S. Y., Yoon, H. I., Jheon, S. H., Sung, S. W., and Hong, Y. C. (2007). Aryl hydrocarbon receptor gene polymorphisms affect lung cancer risk. Lung cancer (Amsterdam, Netherlands) *56*, 9-15.
Kingsley, L. A., Fournier, P. G., Chirgwin, J. M., and Guise, T. A. (2007). Molecular biology of bone metastasis. Molecular cancer therapeutics *6*, 2609-2617.
Klein, A., Wessel, R., Graessmann, M., Jurgens, M., Petersen, I., Schmutzler, R., Niederacher, D., Arnold, N., Meindl, A., Scherneck, S.*, et al.* (2007). Comparison of gene expression data from human and mouse breast cancers: identification of a conserved breast tumor gene set. Int J Cancer *121*, 683-688.
Knudtson, K. L., Auer, H., Brooks, A. I., Griffin, C., Grills, G., Hester, S., Khitrov, G., Lilley, K. S., Massimi, A., Tiesman, J. P., and Viale, A. (2006). The ABRF MARG microarray survey 2005: taking the pulse of the microarray field. J Biomol Tech *17*, 176-186.
Knutsen, H. K., Tasken, K., Eskild, W., Richards, J. S., Kurten, R. C., Torjesen, P. A., Jahnsen, T., Hansson, V., Guerin, S., and Tasken, K. A. (1997). Characterization of the 5'-flanking region of the gene for the cAMP-inducible protein kinase A subunit, RIIbeta, in Sertoli cells. Molecular and cellular endocrinology *129*, 101-114.
Kohonen, T. (1995). Self Organizing Maps, (Berlin: Springer).
Kusuhara, M., Yamaguchi, K., Nagasaki, K., Hayashi, C., Suzaki, A., Hori, S., Handa, S., Nakamura, Y., and Abe, K. (1990). Production of endothelin in human cancer cell lines. Cancer Res *50*, 3257-3261.
Kwoh, C. K., and Ng, P. Y. (2007). Network analysis approach for biology. Cell Mol Life Sci *64*, 1739-1751.
Lander, E. S. (1999). Array of hope. Nature genetics *21*, 3-4.
Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.
Landgrebe, J., Wurst, W., and Welzl, G. (2002). Permutation-validated principal components analysis of microarray data. Genome biology *3*, RESEARCH0019.
Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science *262*, 208-214.
Le Devedec, S. E., Yan, K., de Bont, H., Ghotra, V., Truong, H., Danen, E. H., Verbeek, F., and van de Water, B. (2010). Systems microscopy approaches to understand cancer cell migration and metastasis. Cell Mol Life Sci *67*, 3219-3240.
Lee, F. S., Hagler, J., Chen, Z. J., and Maniatis, T. (1997). Activation of the IkappaB alpha kinase complex by MEKK1, a kinase of the JNK pathway. Cell *88*, 213-222.
Lee, H. C., and Wei, Y. H. (2009). Mitochondrial DNA instability and metabolic shift in human cancers. Int J Mol Sci *10*, 674-701.

Lee, J. S., Chu, I. S., Mikaelyan, A., Calvisi, D. F., Heo, J., Reddy, J. K., and Thorgeirsson, S. S. (2004). Application of comparative functional genomics to identify best-fit mouse models to study human cancer. Nature genetics *36*, 1306-1311.

Lee, M. L., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proceedings of the National Academy of Sciences of the United States of America *97*, 9834-9839.

Lelandais, G., Vincens, P., Badel-Chagnon, A., Vialette, S., Jacq, C., and Hazout, S. (2006). Comparing gene expression networks in a multi-dimensional space to extract similarities and differences between organisms. Bioinformatics (Oxford, England) *22*, 1359-1366.

Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N., and Wasserman, W. W. (2003). Identification of conserved regulatory elements by comparative genome analysis. J Biol *2*, 13.

Lerner, A., Kim, D. H., and Lee, R. (2000). The cAMP signaling pathway as a therapeutic target in lymphoid malignancies. Leukemia & lymphoma *37*, 39-51.

Lerner, L. E., Gribanova, Y. E., Ji, M., Knox, B. E., and Farber, D. B. (2001). Nrl and Sp nuclear proteins mediate transcription of rod-specific cGMP-phosphodiesterase beta-subunit gene: involvement of multiple response elements. J Biol Chem *276*, 34999-35007.

Lerner, L. E., Gribanova, Y. E., Whitaker, L., Knox, B. E., and Farber, D. B. (2002). The rod cGMP-phosphodiesterase beta-subunit promoter is a specific target for Sp4 and is not activated by other Sp proteins or CRX. J Biol Chem *277*, 25877-25883.

Levitzki, A., Rudick, J., Pastan, I., Vass, W. C., and Lowy, D. R. (1986). Adenylate cyclase activity of NIH 3T3 cells morphologically transformed by ras genes. FEBS letters *197*, 134-138.

Li, C., and Hung Wong, W. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome biology *2*, RESEARCH0032.

Li, H., Xuan, J., Wang, Y., and Zhan, M. (2008). Inferring regulatory networks. Front Biosci *13*, 263-275.

Li, Z., Van Calcar, S., Qu, C., Cavenee, W. K., Zhang, M. Q., and Ren, B. (2003). A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. Proceedings of the National Academy of Sciences of the United States of America *100*, 8164-8169.

Liao, B. Y., and Zhang, J. (2006a). Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol *23*, 530-540.

Liao, B. Y., and Zhang, J. (2006b). Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Mol Biol Evol *23*, 1119-1128.

Lin, B., White, J. T., Lu, W., Xie, T., Utleg, A. G., Yan, X., Yi, E. C., Shannon, P., Khrebtukova, I., Lange, P. H.*, et al.* (2005). Evidence for the presence of disease-perturbed networks in prostate cancer cells by genomic and proteomic analyses: a systems approach to disease. Cancer Res *65*, 3081-3091.

Lin, C. S., Chow, S., Lau, A., Tu, R., and Lue, T. F. (2001a). Identification and regulation of human PDE5A gene promoter. Biochem Biophys Res Commun *280*, 684-692.

Lin, C. S., Chow, S., Lau, A., Tu, R., and Lue, T. F. (2001b). Regulation of human PDE5A2 intronic promoter by cAMP and cGMP: identification of a critical Sp1-binding site. Biochem Biophys Res Commun *280*, 693-699.
Lin, K. K., Chudova, D., Hatfield, G. W., Smyth, P., and Andersen, B. (2004). Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. Proceedings of the National Academy of Sciences of the United States of America *101*, 15955-15960.
Lin, P., Chang, H., Tsai, W. T., Wu, M. H., Liao, Y. S., Chen, J. T., and Su, J. M. (2003). Overexpression of aryl hydrocarbon receptor in human lung carcinomas. Toxicologic pathology *31*, 22-30.
Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. Nature genetics *21*, 20-24.
Liu, H., Palmer, D., Jimmo, S. L., Tilley, D. G., Dunkerley, H. A., Pang, S. C., and Maurice, D. H. (2000). Expression of phosphodiesterase 4D (PDE4D) is regulated by both the cyclic AMP-dependent protein kinase and mitogen-activated protein kinase signaling pathways. A potential mechanism allowing for the coordinated regulation of PDE4D activity and expression in cells. J Biol Chem *275*, 26615-26624.
Liu, W. M., Mei, R., Di, X., Ryder, T. B., Hubbell, E., Dee, S., Webster, T. A., Harrington, C. A., Ho, M. H., Baid, J., and Smeekens, S. P. (2002). Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics (Oxford, England) *18*, 1593-1599.
Liu, X., Brutlag, D. L., and Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput, 127-138.
Liu, X., Milo, M., Lawrence, N. D., and Rattray, M. (2005). A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. Bioinformatics (Oxford, England) *21*, 3637-3644.
Locasale, J. W., Cantley, L. C., and Vander Heiden, M. G. (2009). Cancer's insatiable appetite. Nature biotechnology *27*, 916-917.
Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature biotechnology *14*, 1675-1680.
Lockhart, D. J., and Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. Nature *405*, 827-836.
Long, J. R., Egan, K. M., Dunning, L., Shu, X. O., Cai, Q., Cai, H., Dai, Q., Holtzman, J., Gao, Y. T., and Zheng, W. (2006). Population-based case-control study of AhR (aryl hydrocarbon receptor) and CYP1A2 polymorphisms and breast cancer risk. Pharmacogenetics and genomics *16*, 237-243.
Lu, J., Sharma, L. K., and Bai, Y. (2009). Implications of mitochondrial DNA mutations and mitochondrial dysfunction in tumorigenesis. Cell Res *19*, 802-815.
Lu, Z., Xu, S., Joazeiro, C., Cobb, M. H., and Hunter, T. (2002). The PHD domain of MEKK1 acts as an E3 ubiquitin ligase and mediates ubiquitination and degradation of ERK1/2. Molecular cell *9*, 945-956.
MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Paper presented at: Fifth Berkeley Symposium on Mathematical Statistics and Probability.

Malumbres, M., and Barbacid, M. (2003). RAS oncogenes: the first 30 years. Nat Rev Cancer *3*, 459-465.
Mantovani, G., Bondioni, S., Lania, A. G., Rodolfo, M., Peverelli, E., Polentarutti, N., Veliz Rodriguez, T., Ferrero, S., Bosari, S., Beck-Peccoz, P., and Spada, A. (2008). High expression of PKA regulatory subunit 1A protein is related to proliferation of human melanoma cells. Oncogene *27*, 1834-1843.
Mao, D. Y., Watson, J. D., Yan, P. S., Barsyte-Lovejoy, D., Khosravi, F., Wong, W. W., Farnham, P. J., Huang, T. H., and Penn, L. Z. (2003). Analysis of Myc bound loci identified by CpG island arrays shows that Max is essential for Myc-dependent repression. Curr Biol *13*, 882-886.
Marais, R., Light, Y., Mason, C., Paterson, H., Olson, M. F., and Marshall, C. J. (1998). Requirement of Ras-GTP-Raf complexes for activation of Raf-1 by protein kinase C. Science (New York, NY *280*, 109-112.
Marlowe, J. L., and Puga, A. (2005). Aryl hydrocarbon receptor, cell cycle regulation, toxicity, and tumorigenesis. Journal of cellular biochemistry *96*, 1174-1184.
Marton, M. J., DeRisi, J. L., Bennett, H. A., Iyer, V. R., Meyer, M. R., Roberts, C. J., Stoughton, R., Burchard, J., Slade, D., Dai, H.*, et al.* (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. Nature medicine *4*, 1293-1301.
Maser, R. S., Choudhury, B., Campbell, P. J., Feng, B., Wong, K. K., Protopopov, A., O'Neil, J., Gutierrez, A., Ivanova, E., Perna, I.*, et al.* (2007). Chromosomally unstable mouse tumours have genomic alterations similar to diverse human cancers. Nature *447*, 966-971.
Maurer, G. D., Leupold, J. H., Schewe, D. M., Biller, T., Kates, R. E., Hornung, H. M., Lau-Werner, U., Post, S., and Allgayer, H. (2007). Analysis of specific transcriptional regulators as early predictors of independent prognostic relevance in resected colorectal cancer. Clin Cancer Res *13*, 1123-1132.
Mayr, B., and Montminy, M. (2001). Transcriptional regulation by the phosphorylation-dependent factor CREB. Nat Rev Mol Cell Biol *2*, 599-609.
Mazurek, S., and Eigenbrodt, E. (2003). The tumor metabolome. Anticancer research *23*, 1149-1154.
McCue, L. A., Thompson, W., Carmack, C. S., and Lawrence, C. E. (2002). Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res *12*, 1523-1532.
McFate, T., Mohyeldin, A., Lu, H., Thakar, J., Henriques, J., Halim, N. D., Wu, H., Schell, M. J., Tsang, T. M., Teahan, O.*, et al.* (2008). Pyruvate dehydrogenase complex activity controls metabolic and malignant phenotype in cancer cells. J Biol Chem *283*, 22700-22708.
McGuire, A. M., Hughes, J. D., and Church, G. M. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Res *10*, 744-757.
McKay, M. M., and Morrison, D. K. (2007). Integrating signals from RTKs to ERK/MAPK. Oncogene *26*, 3113-3121.
Mei, F. C., Qiao, J., Tsygankova, O. M., Meinkoth, J. L., Quilliam, L. A., and Cheng, X. (2002). Differential signaling of cyclic AMP: opposing effects of exchange protein directly activated by cyclic AMP and cAMP-dependent protein kinase on protein kinase B activation. J Biol Chem *277*, 11497-11504.

Melis, M., Hernandez, J., Siegel, E. M., McLoughlin, J. M., Ly, Q. P., Nair, R. M., Lewis, J. M., Jensen, E. H., Alvarado, M. D., Coppola, D.*, et al.* (2010). Gene expression profiling of colorectal mucinous adenocarcinomas. Diseases of the colon and rectum *53*, 936-943.
Menendez, D., Inga, A., and Resnick, M. A. (2009). The expanding universe of p53 targets. Nat Rev Cancer *9*, 724-737.
Meunier, L., Puiffe, M. L., Le Page, C., Filali-Mouhim, A., Chevrette, M., Tonin, P. N., Provencher, D. M., and Mes-Masson, A. M. (2010). Effect of ovarian cancer ascites on cell migration and gene expression in an epithelial ovarian cancer in vitro model. Translational oncology *3*, 230-238.
Meuwissen, R., and Berns, A. (2005). Mouse models for human lung cancer. Genes Dev *19*, 643-664.
Meyer, T. E., Waeber, G., Lin, J., Beckmann, W., and Habener, J. F. (1993). The promoter of the gene encoding 3',5'-cyclic adenosine monophosphate (cAMP) response element binding protein contains cAMP response elements: evidence for positive autoregulation of gene transcription. Endocrinology *132*, 770-780.
Meyniel, J. P., Cottu, P. H., Decraene, C., Stern, M. H., Couturier, J., Lebigot, I., Nicolas, A., Weber, N., Fourchotte, V., Alran, S.*, et al.* (2010). A genomic and transcriptomic approach for a differential diagnosis between primary and secondary ovarian carcinomas in patients with a previous history of breast cancer. BMC cancer *10*, 222.
Michaels, G. S., Carr, D. B., Askenazi, M., Fuhrman, S., Wen, X., and Somogyi, R. (1998). Cluster analysis and data visualization of large-scale gene expression data. Pacific Symposium on Biocomputing, 42-53.
Mikula, M., Rubel, T., Karczmarski, J., Goryca, K., Dadlez, M., and Ostrowski, J. (2010). Integrating proteomic and transcriptomic high-throughput surveys for search of new biomarkers of colon tumors. Functional & integrative genomics.
Miller, W., Makova, K. D., Nekrutenko, A., and Hardison, R. C. (2004). Comparative genomics. Annual review of genomics and human genetics *5*, 15-56.
Minamino, T., Yujiri, T., Terada, N., Taffet, G. E., Michael, L. H., Johnson, G. L., and Schneider, M. D. (2002). MEKK1 is essential for cardiac hypertrophy and dysfunction induced by Gq. Proceedings of the National Academy of Sciences of the United States of America *99*, 3866-3871.
Mo, X., Kowenz-Leutz, E., Xu, H., and Leutz, A. (2004). Ras induces mediator complex exchange on C/EBP beta. Molecular cell *13*, 241-250.
Mohamed, M. K., Taylor, R. E., Feinstein, D. S., Huang, X., and Pittler, S. J. (1998). Structure and upstream region characterization of the human gene encoding rod photoreceptor cGMP phosphodiesterase alpha-subunit. J Mol Neurosci *10*, 235-250.
Monsieurs, P., Thijs, G., Fadda, A. A., De Keersmaecker, S. C., Vanderleyden, J., De Moor, B., and Marchal, K. (2006). More robust detection of motifs in coexpressed genes by using phylogenetic information. BMC Bioinformatics *7*, 160.
Morey, J. S., Ryan, J. C., and Van Dolah, F. M. (2006). Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. Biological procedures online *8*, 175-193.
Nagaraju, G., and Scully, R. (2007). Minding the gap: the underground functions of BRCA1 and BRCA2 at stalled replication forks. DNA repair *6*, 1018-1031.
Nakamura, Y., Yujiri, T., Nawata, R., Tagami, K., and Tanizawa, Y. (2005). MEK kinase 1 is essential for Bcr-Abl-induced STAT3 and self-renewal activity in embryonic stem cells. Oncogene *24*, 7592-7598.

Nebl, G., Mermod, N., and Cato, A. C. (1994). Post-transcriptional down-regulation of expression of transcription factor NF1 by Ha-ras oncogene. J Biol Chem *269*, 7371-7378.
Negus, R. P., Stamp, G. W., Hadley, J., and Balkwill, F. R. (1997). Quantitative assessment of the leukocyte infiltrate in ovarian cancer and its relationship to the expression of C-C chemokines. Am J Pathol *150*, 1723-1734.
Negus, R. P., Stamp, G. W., Relf, M. G., Burke, F., Malik, S. T., Bernasconi, S., Allavena, P., Sozzani, S., Mantovani, A., and Balkwill, F. R. (1995). The detection and localization of monocyte chemoattractant protein-1 (MCP-1) in human ovarian cancer. J Clin Invest *95*, 2391-2396.
Nesterova, M., Yokozaki, H., McDuffie, E., and Cho-Chung, Y. S. (1996). Overexpression of RII beta regulatory subunit of protein kinase A in human colon carcinoma cell induces growth arrest and phenotypic changes that are abolished by site-directed mutation of RII beta. Eur J Biochem *235*, 486-494.
Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. Psychological methods *5*, 241-301.
Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J. M., and Pascual-Montano, A. (2009). GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. Nucleic Acids Res *37*, W317-322.
O'Hagan, R. C., Chang, S., Maser, R. S., Mohan, R., Artandi, S. E., Chin, L., and DePinho, R. A. (2002). Telomere dysfunction provokes regional amplification and deletion in cancer genomes. Cancer cell *2*, 149-155.
Oda, K., Matsuoka, Y., Funahashi, A., and Kitano, H. (2005). A comprehensive pathway map of epidermal growth factor receptor signaling. Molecular systems biology *1*, 2005 0010.
Orntoft, T. F., Thykjaer, T., Waldman, F. M., Wolf, H., and Celis, J. E. (2002). Genome-wide study of gene copy numbers, transcripts, and protein levels in pairs of non-invasive and invasive human transitional cell carcinomas. Mol Cell Proteomics *1*, 37-45.
Ovaska, K., Laakso, M., Haapa-Paananen, S., Louhimo, R., Chen, P., Aittomaki, V., Valo, E., Nunez-Fontarnau, J., Rantanen, V., Karinen, S.*, et al.* (2010). Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome Med *2*, 65.
Pantel, K., and Brakenhoff, R. H. (2004). Dissecting the metastatic cascade. Nat Rev Cancer *4*, 448-456.
Pavlidis, P., and Noble, W. S. (2001). Analysis of strain and regional variation in gene expression in mouse brain. Genome biology *2*, RESEARCH0042.
Peterson, L. E. (2003). Partitioning large-sample microarray-based gene expression profiles using principal components analysis. Computer methods and programs in biomedicine *70*, 107-119.
Piccoli, C., Scacco, S., Bellomo, F., Signorile, A., Iuso, A., Boffoli, D., Scrima, R., Capitanio, N., and Papa, S. (2006). cAMP controls oxygen metabolism in mammalian cells. FEBS Lett *580*, 4539-4543.
Pocar, P., Fischer, B., Klonisch, T., and Hombach-Klonisch, S. (2005). Molecular interactions of the aryl hydrocarbon receptor and its biological and toxicological relevance for reproduction. Reproduction (Cambridge, England) *129*, 379-389.
Pore, N., Liu, S., Shu, H. K., Li, B., Haas-Kogan, D., Stokoe, D., Milanini-Mongiat, J., Pages, G., O'Rourke, D. M., Bernhard, E., and Maity, A. (2004). Sp1 is involved

in Akt-mediated induction of VEGF expression through an HIF-1-independent mechanism. Mol Biol Cell *15*, 4841-4853.
Pratap, J., Galindo, M., Zaidi, S. K., Vradii, D., Bhat, B. M., Robinson, J. A., Choi, J. Y., Komori, T., Stein, J. L., Lian, J. B., *et al.* (2003). Cell growth regulatory role of Runx2 during proliferative expansion of preosteoblasts. Cancer Res *63*, 5357-5362.
Pratap, J., Lian, J. B., Javed, A., Barnes, G. L., van Wijnen, A. J., Stein, J. L., and Stein, G. S. (2006). Regulatory roles of Runx2 in metastatic tumor and cancer cell interactions with bone. Cancer metastasis reviews *25*, 589-600.
Pratap, J., Wixted, J. J., Gaur, T., Zaidi, S. K., Dobson, J., Gokul, K. D., Hussain, S., van Wijnen, A. J., Stein, J. L., Stein, G. S., and Lian, J. B. (2008). Runx2 transcriptional activation of Indian Hedgehog and a downstream bone metastatic pathway in breast cancer cells. Cancer Res *68*, 7795-7802.
Pritchard, C., Mecham, B., Dumpit, R., Coleman, I., Bhattacharjee, M., Chen, Q., Sikes, R. A., and Nelson, P. S. (2009). Conserved gene expression programs integrate mammalian prostate development and tumorigenesis. Cancer Res *69*, 1739-1747.
Pritchard, C. C., Hsu, L., Delrow, J., and Nelson, P. S. (2001). Project normal: defining normal variance in mouse gene expression. Proceedings of the National Academy of Sciences of the United States of America *98*, 13266-13271.
Raha, S., Myint, A. T., Johnstone, L., and Robinson, B. H. (2002). Control of oxygen free radical formation from mitochondrial complex I: roles for protein kinase A and pyruvate dehydrogenase kinase. Free Radic Biol Med *32*, 421-430.
Rajeevan, M. S., Ranamukhaarachchi, D. G., Vernon, S. D., and Unger, E. R. (2001). Use of real-time quantitative PCR to validate the results of cDNA array and differential display PCR technologies. Methods (San Diego, Calif *25*, 443-451.
Ramanathan, A., Wang, C., and Schreiber, S. L. (2005). Perturbational profiling of a cell-line model of tumorigenesis by using metabolic measurements. Proceedings of the National Academy of Sciences of the United States of America *102*, 5992-5997.
Ramaswamy, S., Ross, K. N., Lander, E. S., and Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. Nature genetics *33*, 49-54.
Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. Pacific Symposium on Biocomputing, 455-466.
Rhodes, D. R., and Chinnaiyan, A. M. (2005). Integrative analysis of the cancer transcriptome. Nature genetics *37 Suppl*, S31-37.
Ringner, M. (2008). What is principal component analysis? Nature biotechnology *26*, 303-304.
Robinson-White, A. J., Bossis, I., Hsiao, H. P., Nesterova, M., Leitner, W. W., and Stratakis, C. A. (2009). 8-Cl-adenosine inhibits proliferation and causes apoptosis in B-lymphocytes via protein kinase A-dependent and independent effects: implications for treatment of Carney complex-associated tumors. J Clin Endocrinol Metab *94*, 4061-4069.
Rosenow, C., Vailaya, A., Kuchinsky, A. J., and Middleton, F. A. (2005). Combining expression and genotyping analysis in neuropsychiatric research. Agilent Technologies.
Roses, A. D. (2000). Pharmacogenetics and the practice of medicine. Nature *405*, 857-865.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., *et al.* (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet *24*, 227-235.
Saal, L. H., Johansson, P., Holm, K., Gruvberger-Saal, S. K., She, Q. B., Maurer, M., Koujak, S., Ferrando, A. A., Malmstrom, P., Memeo, L., *et al.* (2007). Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. Proceedings of the National Academy of Sciences of the United States of America *104*, 7564-7569.
Saghir, F. S., Rose, I. M., Dali, A. Z., Shamsuddin, Z., Jamal, A. R., and Mokhtar, N. M. (2010). Gene expression profiling and cancer-related pathways in type I endometrial carcinoma. Int J Gynecol Cancer *20*, 724-731.
Salas, T. R., Reddy, S. A., Clifford, J. L., Davis, R. J., Kikuchi, A., Lippman, S. M., and Menter, D. G. (2003). Alleviating the suppression of glycogen synthase kinase-3beta by Akt leads to the phosphorylation of cAMP-response element-binding protein and its transactivation in intact cell nuclei. J Biol Chem *278*, 41338-41346.
Salton, M., Lerenthal, Y., Wang, S. Y., Chen, D. J., and Shiloh, Y. (2010). Involvement of matrin 3 and SFPQ/NONO in the DNA damage response. Cell cycle (Georgetown, Tex *9*.
Santini, S., Boore, J. L., and Meyer, A. (2003). Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. Genome Res *13*, 1111-1122.
Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (New York, NY *270*, 467-470.
Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T., *et al.* (2000). A gene expression database for the molecular pharmacology of cancer. Nature genetics *24*, 236-244.
Schlezinger, J. J., Liu, D., Farago, M., Seldin, D. C., Belguise, K., Sonenshein, G. E., and Sherr, D. H. (2006). A role for the aryl hydrocarbon receptor in mammary gland tumorigenesis. Biological chemistry *387*, 1175-1187.
Sears, R. C. (2004). The life cycle of C-myc: from synthesis to degradation. Cell cycle (Georgetown, Tex *3*, 1133-1137.
Segal, E., Friedman, N., Kaminski, N., Regev, A., and Koller, D. (2005). From signatures to models: understanding cancer using microarrays. Nature genetics *37 Suppl*, S38-45.
Sell, S. (2003). Mouse models to study the interaction of risk factors for human liver cancer. Cancer Res *63*, 7553-7562.
Shell, S. A., Fix, C., Olejniczak, D., Gram-Humphrey, N., and Walker, W. H. (2002). Regulation of cyclic adenosine 3',5'-monophosphate response element binding protein (CREB) expression by Sp1 in the mammalian testis. Biology of reproduction *66*, 659-666.
Sherlock, G. (2001). Analysis of large-scale gene expression data. Brief Bioinform *2*, 350-362.
Simonnet, H., Demont, J., Pfeiffer, K., Guenaneche, L., Bouvier, R., Brandt, U., Schagger, H., and Godinot, C. (2003). Mitochondrial complex I is deficient in renal oncocytomas. Carcinogenesis *24*, 1461-1466.
Singh, I. S., Luo, Z., Kozlowski, M. T., and Erlichman, J. (1994). Association of USF and c-Myc with a helix-loop-helix-consensus motif in the core promoter of the murine type II beta regulatory subunit gene of cyclic adenosine 3', 5'-monophosphate-dependent protein kinase. Mol Endocrinol *8*, 1163-1174.

Siu, Y. T., Ching, Y. P., and Jin, D. Y. (2008). Activation of TORC1 transcriptional coactivator through MEKK1-induced phosphorylation. Mol Biol Cell *19*, 4750-4761.

Solberg, R., Sandberg, M., Natarajan, V., Torjesen, P. A., Hansson, V., Jahnsen, T., and Tasken, K. (1997). The human gene for the regulatory subunit RI alpha of cyclic adenosine 3', 5'-monophosphate-dependent protein kinase: two distinct promoters provide differential regulation of alternately spliced messenger ribonucleic acids. Endocrinology *138*, 169-181.

Somyajit, K., Subramanya, S., and Nagaraju, G. (2010). RAD51C: a novel cancer susceptibility gene is linked to Fanconi anemia and breast cancer. Carcinogenesis *31*, 2031-2038.

Spina, A. M., Chiosi, E., Naviglio, S., Valente, F., Marchese, M., Furgi, A., Metafora, S., and Illiano, G. (1993). ras oncogene-induced transformation of a rat seminal vesicle epithelial cell line produces a marked increase of adenylate cyclase and protein kinase C activities. FEBS letters *331*, 150-154.

Srinivasan, B. S., Shah, N. H., Flannick, J. A., Abeliuk, E., Novak, A. F., and Batzoglou, S. (2007). Current progress in network research: toward reference networks for key model organisms. Brief Bioinform *8*, 318-332.

Strand, A. D., Aragaki, A. K., Baquet, Z. C., Hodges, A., Cunningham, P., Holmans, P., Jones, K. R., Jones, L., Kooperberg, C., and Olson, J. M. (2007). Conservation of regional gene expression in mouse and human brain. PLoS Genet *3*, e59.

Strausberg, R. L. (2001). The Cancer Genome Anatomy Project: new resources for reading the molecular signatures of cancer. J Pathol *195*, 31-40.

Strausberg, R. L., Buetow, K. H., Emmert-Buck, M. R., and Klausner, R. D. (2000). The cancer genome anatomy project: building an annotated gene index. Trends Genet *16*, 103-106.

Streb, J. W., and Miano, J. M. (2005). AKAP12alpha, an atypical serum response factor-dependent target gene. J Biol Chem *280*, 4125-4134.

Stuart, R. O., Bush, K. T., and Nigam, S. K. (2001). Changes in global gene expression patterns during development and maturation of the rat kidney. Proceedings of the National Academy of Sciences of the United States of America *98*, 5649-5654.

Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A.*, et al.* (2002). Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences of the United States of America *99*, 4465-4470.

Sumiyama, K., Kim, C. B., and Ruddle, F. H. (2001). An efficient cis-element discovery method using multiple sequence comparisons based on evolutionary relationships. Genomics *71*, 260-262.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J. J., Ladd-Acosta, C., Mesirov, J., Golub, T. R., and Jacks, T. (2005). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nature genetics *37*, 48-55.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the United States of America *96*, 2907-2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. Nat Genet *22*, 281-285.
Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2002). A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol *9*, 447-464.
Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J.*, et al.* (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol *23*, 137-144.
Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. (1999). Analysis of gene expression data using self-organizing maps. FEBS letters *451*, 142-146.
Torras-Llort, M., and Azorin, F. (2003). Functional characterization of the human phosphodiesterase 7A1 promoter. Biochem J *373*, 835-843.
Tortora, G., and Ciardiello, F. (2002). Protein kinase A as target for novel integrated strategies of cancer therapy. Ann N Y Acad Sci *968*, 139-147.
Treisman, R. (1996). Regulation of transcription by MAP kinase cascades. Current opinion in cell biology *8*, 205-215.
Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America *98*, 5116-5121.
van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T.*, et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530-536.
van Helden, J., Andre, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol *281*, 827-842.
Van Hellemont, R., Monsieurs, P., Thijs, G., de Moor, B., Van de Peer, Y., and Marchal, K. (2005). A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biol *6*, R113.
Vesanto, J. (1999). SOM-Based data visualization methods. Intelligent Data Analysis journal.
Vesanto, J., and Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Trans Neural Netw.
Vogel, K. S., Klesse, L. J., Velasco-Miguel, S., Meyers, K., Rushing, E. J., and Parada, L. F. (1999). Mouse tumor model for neurofibromatosis type 1. Science (New York, NY *286*, 2176-2179.
Vogelstein, B., and Kinzler, K. W. (2004). Cancer genes and the pathways they control. Nature medicine *10*, 789-799.
Vogt, P. K., and Bos, T. J. (1990). jun: oncogene and transcription factor. Advances in cancer research *55*, 1-35.
Vohradsky, J., Li, X. M., and Thompson, C. J. (1997). Identification of procaryotic developmental stages by statistical analyses of two-dimensional gel patterns. Electrophoresis *18*, 1418-1428.
Waghray, A., Feroze, F., Schober, M. S., Yao, F., Wood, C., Puravs, E., Krause, M., Hanash, S., and Chen, Y. Q. (2001). Identification of androgen-regulated genes in the prostate cancer cell line LNCaP by serial analysis of gene expression and proteomic analysis. Proteomics *1*, 1327-1338.

Walker, W. H., and Habener, J. F. (1996). Role of transcription factors CREB and CREM in cAMP-regulated transcription during spermatogenesis. Trends in endocrinology and metabolism: TEM *7*, 133-138.
Wang, D., Deng, C., Bugaj-Gaweda, B., Kwan, M., Gunwaldsen, C., Leonard, C., Xin, X., Hu, Y., Unterbeck, A., and De Vivo, M. (2003). Cloning and characterization of novel PDE4D isoforms PDE4D6 and PDE4D7. Cellular signaling *15*, 883-891.
Wang, H., Huang, S., Shou, J., Su, E. W., Onyia, J. E., Liao, B., and Li, S. (2006). Comparative analysis and integrative classification of NCI60 cell lines and primary tumors using gene expression profiling data. BMC genomics *7*, 166.
Wang, H., Yang, J., Wang, W., and Yu, P. S. (2002). Clustering by Pattern Similarity in Large Data Sets. In Proc. ACM SIGMOD Conference, (Madison).
Warburg, O. (1956). On the origin of cancer cells. Science (New York, NY *123*, 309-314.
Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W., and Lawrence, C. E. (2000). Human-mouse genome comparisons to locate regulatory sites. Nat Genet *26*, 225-228.
Waterston, R. H., Lander, E. S., and Sulston, J. E. (2002). On the sequencing of the human genome. Proceedings of the National Academy of Sciences of the United States of America *99*, 3712-3716.
Wellmann, A., Thieblemont, C., Pittaluga, S., Sakai, A., Jaffe, E. S., Siebert, P., and Raffeld, M. (2000). Detection of differentially expressed genes in lymphomas using cDNA arrays: identification of clusterin as a new diagnostic marker for anaplastic large-cell lymphomas. Blood *96*, 398-404.
Wells, A. (1999). EGF receptor. The international journal of biochemistry & cell biology *31*, 637-643.
Witowsky, J. A., and Johnson, G. L. (2003). Ubiquitylation of MEKK1 inhibits its phosphorylation of MKK1 and MKK4 and activation of the ERK1/2 and JNK pathways. J Biol Chem *278*, 1403-1406.
Woelfle, U., Cloos, J., Sauter, G., Riethdorf, L., Janicke, F., van Diest, P., Brakenhoff, R., and Pantel, K. (2003). Molecular signature associated with bone marrow micrometastasis in human breast cancer. Cancer Res *63*, 5679-5684.
Wolfsberg, T. G., Gabrielian, A. E., Campbell, M. J., Cho, R. J., Spouge, J. L., and Landsman, D. (1999). Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae. Genome Res *9*, 775-792.
Wong, W. K., Ou, X. M., Chen, K., and Shih, J. C. (2002). Activation of human monoamine oxidase B gene expression by a protein kinase C MAPK signal transduction pathway involves c-Jun and Egr-1. J Biol Chem *277*, 22222-22230.
Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K.*, et al.* (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol *3*, e7.
Workman, C. T., and Stormo, G. D. (2000). ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pac Symp Biocomput, 467-478.
Wu, F., Dassopoulos, T., Cope, L., Maitra, A., Brant, S. R., Harris, M. L., Bayless, T. M., Parmigiani, G., and Chakravarti, S. (2007). Genome-wide gene expression differences in Crohn's disease and ulcerative colitis from endoscopic pinch biopsies: insights into distinctive pathogenesis. Inflammatory bowel diseases *13*, 807-821.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T. P., and Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. Nat Methods *6*, 75-77.
www.ebi.ac.uk/intact.
Xia, Y., Makris, C., Su, B., Li, E., Yang, J., Nemerow, G. R., and Karin, M. (2000). MEK kinase 1 is critically required for c-Jun N-terminal kinase activation by proinflammatory stimuli and growth factor-induced cell migration. Proceedings of the National Academy of Sciences of the United States of America *97*, 5243-5248.
Xia, Y., Wu, Z., Su, B., Murray, B., and Karin, M. (1998). JNKK1 organizes a MAP kinase module through specific and sequential interactions with upstream and downstream components mediated by its amino-terminal extension. Genes Dev *12*, 3369-3381.
Yamaguchi, H., Kojima, T., Ito, T., Kimura, Y., Imamura, M., Son, S., Koizumi, J., Murata, M., Nagayama, M., Nobuoka, T.*, et al.* (2010). Transcriptional control of tight junction proteins via a protein kinase C signal pathway in human telomerase reverse transcriptase-transfected human pancreatic duct epithelial cells. Am J Pathol *177*, 698-712.
Yamamoto, F., and Perucho, M. (1984). Activation of a human c-K-ras oncogene. Nucleic Acids Res *12*, 8873-8885.
Yang, j. (2002). Capturing subspace correlation in a large data set. Paper presented at: ICDE Conference.
Yarden, Y., and Sliwkowski, M. X. (2001). Untangling the ErbB signaling network. Nat Rev Mol Cell Biol *2*, 127-137.
Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics (Oxford, England) *17*, 977-987.
Young, D. W., Hassan, M. Q., Yang, X. Q., Galindo, M., Javed, A., Zaidi, S. K., Furcinitti, P., Lapointe, D., Montecino, M., Lian, J. B.*, et al.* (2007). Mitotic retention of gene expression patterns by the cell fate-determining transcription factor Runx2. Proceedings of the National Academy of Sciences of the United States of America *104*, 3189-3194.
Yujiri, T., Sather, S., Fanger, G. R., and Johnson, G. L. (1998). Role of MEKK1 in cell survival and activation of JNK and ERK pathways defined by targeted gene disruption. Science (New York, NY *282*, 1911-1914.
Zaidi, S. K., Pande, S., Pratap, J., Gaur, T., Grigoriu, S., Ali, S. A., Stein, J. L., Lian, J. B., van Wijnen, A. J., and Stein, G. S. (2007). Runx2 deficiency and defective subnuclear targeting bypass senescence to promote immortalization and tumorigenic potential. Proceedings of the National Academy of Sciences of the United States of America *104*, 19861-19866.
Zhang, M. Q. (1999). Large-scale gene expression data analysis: a new challenge to computational biologists. Genome research *9*, 681-688.
Zhang, Q., Ushijima, R., Kawai, T., and Tanaka, H. (2004). Which to use? - microarray data analysis in input and output data processing. Chem-Bio Informatics J *4*, 56–72.
Zhao, F., Xuan, Z., Liu, L., and Zhang, M. Q. (2005). TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. Nucleic Acids Res *33*, D103-107.
Zhou, X. J., and Gibson, G. (2004). Cross-species comparison of genome-wide expression patterns. Genome biology *5*, 232.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. Genes Dev *21*, 1010-1024.
Zucchini, C., Rocchi, A., Manara, M. C., De Sanctis, P., Capanni, C., Bianchini, M., Carinci, P., Scotlandi, K., and Valvassori, L. (2008). Apoptotic genes as potential markers of metastatic phenotype in human osteosarcoma cell lines. International journal of oncology *32*, 17-31.